Dirichlet Mixtures in Text Modeling

Mikio Yamamoto and Kugatsu Sadamitsu

CS Technical report CS-TR-05-1
University of Tsukuba

May 30, 2005

Abstract

Word rates in text vary according to global factors such as genre, topic, author, and expected readership (Church and Gale 1995). Models that summarize such global factors in text or at the document level, are called 'text models.' A finite mixture of Dirichlet distribution (Dirichlet Mixture or DM for short) was investigated as a new text model. When parameters of a multinomial are drawn from a DM, the compound for discrete outcomes is a finite mixture of the Dirichlet-multinomial. A Dirichlet multinomial can be regarded as a multivariate version of the Poisson mixture, a reliable univariate model for global factors (Church and Gale 1995). In the present paper, the DM and its compounds are introduced, with parameter estimation methods derived from Minka's fixed-point methods (Minka 2003) and the EM algorithm. The method can estimate a considerable number of parameters of a large DM, i.e., a few hundred thousand parameters. After discussion of the relationships within the DM — probabilistic latent semantic analysis (PLSA) (Hofmann 1999), the mixture of unigrams (Nigam et al. 2000), and latent Dirichlet allocation (LDA) (Blei et al. 2001, 2003) — the products of statistical language modeling applications are discussed and their performance in perplexity compared. The DM model achieves the lowest perplexity level despite its unitopic nature.

1 Introduction

Word rates in text vary according to global factors such as genre, topic, author, and expected readership. Church and Gale (1995) examined the Brown corpus and showed that the English word 'said' occurs with high frequency in the press and fiction, but relatively infrequently in the hobby and learned genres. This observation is basic to their model for word rate variation. Rosenfeld (1999) wrote that the occurrence of the word 'winter' in a document is proportional to the occurrence of the word 'summer' in the same document. This is a basic observation for his trigger models. Church (2001) states that a document including the word 'Noriega' has a high probability for another occurrence of 'Noriega' in the same document, an observation

basic to his adaptation model. In the present paper, an attempt is made to model these global factors with a new generative text model (Dirichlet Mixture or DM for short), and to apply it to improve conventional $n$gram language models that reveal only local interdependency among words.

It is well known that global factors are important to language modeling in three language processing research communities involved in natural language processing, speech processing, and neural networks. In the natural language processing community, Church and Gale (1995) proposed that word rate distribution can be explained by the Poisson mixture — an infinite Poisson mixture model, in which the Poisson parameter varies over the factors of a density function. For example, they demonstrated that an empirical word rate variation closely fits a special case of the Poisson mixture, the negative binomial where the density function assumes a gamma distribution. However, because the Poisson mixture is univariate, it is difficult to manage word rates for every word simultaneously.

Researchers in the speech processing community have proposed and tested a great number of multivariate models — cache models, trigger models and topic-based models — to capture distant word dependency in the past two decades. Iyer and Ostendorf (1999) proposed an m-component mixture of unigram models with a parameter estimation method using the EM algorithm. Each unigram model for the mixture corresponds to a different topic and yields word rates for that topic. Using topic-based finite mixture models, language models can be greatly improved in perplexity, though parameter estimation for this model tends to overfit training data.

Recently, 'generative text models' such as latent Dirichlet allocation (LDA) have attracted people in the neural network community. Using generative text models, the probability for a 'document' rather than simply 'sentences' can be computed. Probability computation in these models takes advantage of prior distribution of word rate variability garnered from large document collections. Generative models are statistically well defined and robust for parameter estimation and adaptation because they exploit (hierarchical) Bayesian frameworks, which rely heavily on a prior and posterior distribution of word rates.

In this paper, a new generative text model was investigated, which unifies the following concepts developed by the three communities related to language processing:

(1) summary of word rate variability as a stochastic distribution

(2) finite topic mixture models of multivariate distributions

(3) generative text models based on a (hierarchical) Bayesian framework

Based on (1) and (2), it was assumed that word rate variability can be modeled with a finite mixture of Dirichlet distributions. Finite mixtures encapsulate rough topic structures, and each Dirichlet distribution yields clear word rate variability within each topic for all words simultaneously. From (3), a robust model is built adopting a Baysian framework, employing a prior and a posterior distribution to estimate DM parameters and to adapt them to the context of thus-processed documents.

In Sections 2 and 3 of the paper, the DM model is described, as are parameters estimation methods, and posterior and predictive distributions of the model. In Section 4, the relationship between DM and other text models is discussed. In Section 5, experimental results of applications for statistical language models are presented. The DM model achieves lower perplexity levels than those employing the mixture of unigrams (MU) and LDA models.

2 The DM and parameter estimation

2.1 The DM and the Polya mixture

The Dirichlet distribution is defined for a random vector, $\boldsymbol{p} = (p_1, p_2 ... p_V)$, on a simplex of V dimensions. Elements of a random vector on a simplex sum to 1. We interpret $\boldsymbol{p}$ as word occurrence probabilities on V words of a vocabulary, so that the Dirichlet distribution models word occurrence probabilities. The density function of the Dirichlet for $\boldsymbol{p}$ is:

$$P_D(\boldsymbol{p}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \prod_{v=1}^{V} p_v^{\alpha_v - 1}, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_V)$ is a parameter vector, $\alpha_v > 0$ and $\alpha = \sum_{v=1}^{V} \alpha_v$.

The Dirichlet mixture distribution (Sjölander et al. 1996) with M components is defined as the following:

$$\begin{aligned} P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) &= \sum_{m=1}^{M} \lambda_m P_D(\boldsymbol{p}; \boldsymbol{\alpha}_m) \\ &= \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\prod_{v=1}^{V} \Gamma(\alpha_{mv})} \prod_{v=1}^{V} p_v^{\alpha_{mv} - 1}, \end{aligned} \tag{2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_M)$ is a weight vector for each component Dirichlet distribution and $\alpha_m = \sum_v \alpha_{mv}$.

When the random vector $\boldsymbol{p}$ as parameters of a multinomial is drawn from the DM, the compound distribution for discrete outcomes $\boldsymbol{y} = (y_1, y_2, ..., y_V)$ is:

$$\begin{aligned} P_{PM}(\boldsymbol{y}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) &= \int P_{Mul}(\boldsymbol{y}|\boldsymbol{p}) P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) d\boldsymbol{p} \\ &= \sum_{m=1}^{M} \lambda_m \int P_{Mul}(\boldsymbol{y}|\boldsymbol{p}) P_D(\boldsymbol{p}; \boldsymbol{\alpha}_m) d\boldsymbol{p} \\ &= \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + y)} \prod_{v=1}^{V} \frac{\Gamma(y_v + \alpha_{mv})}{\Gamma(\alpha_{mv})}, \end{aligned} \tag{3}$$

where

$$y = \sum_v y_v.$$

Each y_v means occurrence frequency of the v-th word in a document. This distribution is called the Dirichlet-multinomial mixture or the Polya mixture distribution. The Polya mixture is used to estimate parameters for the DM, $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$.

2.2 Parameter estimation

In this subsection, methods for estimating parameters were introduced for the DM with a maximum likelihood estimator of the Polya mixture. The estimating methods are based on Minka's estimation methods for a Dirichlet distribution (Minka 2003) and the EM algorithm (Dempster et al. 1977).

Given the i-th datum or the i-th training document, outcomes can be determined for words $\boldsymbol{y}_i = (y_{i1}, y_{i2}...y_{iV})$. For N training documents, the log likelihood function for the training documents $\boldsymbol{D} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$ is:

$$\mathcal{L}(\boldsymbol{D}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) = \sum_{i=1}^{N} \log P_{PM}(\boldsymbol{y}_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M).$$

The $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ that maximize the above likelihood function are also DM parameters.

Assuming $\boldsymbol{Z} = (z_1, z_2, ..., z_N)$ and z_i is a hidden variable that denotes a component generating the i-th document, the log likelihood for the complete data is:

$$\mathcal{L}(\boldsymbol{D}, \boldsymbol{Z}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) = \sum_{i=1}^{N} \log P(\boldsymbol{y}_i, z_i; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M).$$

The Q-function for the EM algorithm or the conditional expectation of the above log likelihood is:

$$\begin{aligned} Q(\theta|\bar{\theta}) &= \sum_i \sum_m P_{im} \log P(\boldsymbol{y}_i, z_i = m; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M) \\ &= \sum_i \sum_m P_{im} \log \lambda_m + \sum_i \sum_m P_{im} \log \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + y_i)} \prod_{v=1}^{V} \frac{\Gamma(y_{iv} + \alpha_{mv})}{\Gamma(\alpha_{mv})}, \end{aligned} \tag{4}$$

where

$$\begin{aligned} P_{im} &= P(z_i = m|\boldsymbol{y}_i; \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}_1^M) \\ y_i &= \sum_v y_{iv}. \end{aligned}$$

$\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\alpha}}_1^M$ are current values of parameters. The first term of (4) can be maximized via the following update formula for $\boldsymbol{\lambda}$.

$$\lambda_m \propto \sum_i P_{im} \tag{5}$$

The second term of (4) can be maximized via the following update formula for $\boldsymbol{\alpha}$. $\Psi(x)$ is the digamma function. The update formula is derived from the Minka's fixed-point iteration (Minka 2003) and the EM algorithm (see Appendix A).

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_i P_{im} \{\Psi(y_{ik} + \bar{\alpha}_{mv}) - \Psi(\bar{\alpha}_{mv})\}}{\sum_i P_{im} \{\Psi(y_i + \bar{\alpha}_m) - \Psi(\bar{\alpha}_m)\}} \tag{6}$$

If the leaving-one-out (LOO) likelihood is designated as the function to be maximized, a faster update formula is obtained. The following update formula is based on Minka's iteration for the LOO likelihood (Minka 2003) and the EM algorithm (see Appendix B):

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_i P_{im} \{y_{iv}/(y_{iv} - 1 + \bar{\alpha}_{mv})\}}{\sum_i P_{im} \{y_i/(y_i - 1 + \bar{\alpha}_m)\}} \tag{7}$$

This LOO update function is used to estimate the DM parameters in all following experiments.

3 Inference

3.1 A posterior and predictive distribution

The DM model with parameters estimated using the above methods is regarded as a prior for the distribution of word occurrence probabilities. In this section, a method is described for computing a posterior and expectations for word occurrence probability, given a word history or a document. The following formula is a posterior distribution for word occurrence probability given the data history $\boldsymbol{y} = (y_1, y_2, ..., y_V)$, assuming a multinomial distribution for count data $\boldsymbol{y}$, with parameter $\boldsymbol{p}$ distributed according to a DM with parameter $\boldsymbol{\alpha}$ as a prior:

$$\begin{aligned} P(\boldsymbol{p}|\boldsymbol{y}) &= \frac{P(\boldsymbol{y}|\boldsymbol{p})P(\boldsymbol{p})}{\int P(\boldsymbol{y}|\boldsymbol{p})P(\boldsymbol{p})d\boldsymbol{p}} \\ &= \frac{P_{Mul}(\boldsymbol{y}|\boldsymbol{p})P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M)}{\int P_{Mul}(\boldsymbol{y}|\boldsymbol{p})P_{DM}(\boldsymbol{p}; \boldsymbol{\lambda}, \boldsymbol{\alpha}_1^M)d\boldsymbol{p}} \\ &= \frac{\sum_{m=1}^M B_m \prod_v p_v^{\alpha_{mv}+y_v-1}}{\sum_{m=1}^M C_m}, \end{aligned} \tag{8}$$

where

$$
\begin{aligned}
B_m &= \lambda_m \frac{\Gamma(\alpha_m)}{\prod_{v=1}^{V} \Gamma(\alpha_{mv})}, \\
C_m &= B_m \frac{\prod_{v=1}^{V} \Gamma(\alpha_{mv} + y_v)}{\Gamma(\alpha_m + y)}, \\
\alpha_m &= \sum_{v=1}^{V} \alpha_{mv}, \\
y &= \sum_{v=1}^{V} y_v.
\end{aligned}
$$

Expectation of occurrence probability of the w-th word in a vocabulary, $P(w^*|\boldsymbol{y})$, is:

$$
\begin{aligned}
P(w^*|\boldsymbol{y}) &= \int p_w P(\boldsymbol{p}|\boldsymbol{y}) d\boldsymbol{p} \\
&= \frac{\sum_{m=1}^{M} B_m \int \prod_{v=1}^{V} p_v^{\alpha_{mv}+y_v+\delta(v-w)-1} d\boldsymbol{p}}{\sum_{m=1}^{M} C_m} \\
&= \frac{\sum_{m=1}^{M} B_m \prod_{v=1}^{V} \frac{\Gamma\{\alpha_{mv}+y_v+\delta(v-w)\}}{\Gamma(\alpha_m+y+1)}}{\sum_{m=1}^{M} C_m} \\
&= \frac{\sum_{m=1}^{M} C_m \frac{\alpha_{mw}+y_w}{\alpha_m+y}}{\sum_{m=1}^{M} C_m}
\end{aligned}
\tag{9}
$$

where $\delta(k)$ is Kronecker's delta,

$$
\delta(k) = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{others .} \end{cases}
$$

In contrast to LDA, DM has a closed formula for computing expectation of word occurrence probability.

3.2 Model averaging

In the experiment section (Section 5), it is demonstrated that a statistical language model using DM outperforms other models with a fewer components, but that performance does not rise in proportion to the number of components. This problem reflects the overfitting nature of DM models.

Avoiding the overfitting problem, a simple model averaging method is adopted, which computes a predictive distribution as a mean for each prediction of DM with a different number of components. It is assumed that there are N different DM models, and that $P^i(w^*|\boldsymbol{y}), i = 1, 2...N$ is a predictive probability for the word w. The following averaging

equation is referred to as method 1, in which evidence probability for a history is regarded as credit weight for each model:

$$P_{ma1}(w^*|\boldsymbol{y}) = \sum_i \frac{P^i_{PM}(\boldsymbol{y};\boldsymbol{\lambda},\boldsymbol{\alpha})}{\sum_j P^j_{PM}(\boldsymbol{y};\boldsymbol{\lambda},\boldsymbol{\alpha})} P^i(w^*|\boldsymbol{y}) \tag{10}$$

Method 2 is a simpler method, averaging predictive probabilities with a simple arithmetic mean:

$$P_{ma2}(w^*|\boldsymbol{y}) = \frac{1}{N}\sum_i P^i(w^*|\boldsymbol{y}) \tag{11}$$

4 Relationship with other topic-based models

The relationship of the DM with the other topic-based models is demonstrated using graphical representation of models. The following are evidence probabilities for data $\boldsymbol{y} = (y_1, y_2, ..., y_V)$ in each topic-based model.

Mixture of unigrams: $P(\boldsymbol{y}) = \sum_z p(z) P_{Mul}(\boldsymbol{y}|z)$

LDA: $P(\boldsymbol{y}) = \int P_D(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_v P(w_v|\boldsymbol{\theta})^{y_v} d\boldsymbol{\theta}$, where $P(w_v|\boldsymbol{\theta}) = \sum_z p(z|\boldsymbol{\theta}) p(w_v|z)$

DM: $P(\boldsymbol{y}) = P_{PM}(\boldsymbol{y})$

Z's of the mixture of unigram models and LDA are latent variables representing topics. $\boldsymbol{\theta}$ of LDA is weights for each unigram modeled with the Dirichlet distribution $P_D(\boldsymbol{\theta};\boldsymbol{\alpha})$. Evidence probability for DM is the Polya distribution described in Sec. 2.1.

Figure 1 is a graphical representation of four models including a probabilistic LSA (PLSA). Outer and inner squares represent N documents (d) and L words (w) in each document, respectively. Circles are random variables and double circles are model parameters. Arrows indicate conditional dependent relationships between variables.

The simplest model is MU. In this model, it is assumed that a document is generated from just one topic — the first chosen topic z, is used to generate all words in the document. The PLSA model can relax this assumption. Under PLSA, it is possible that each word in a document is generated with different topics. In the result, a document is assumed to have multiple topics, a realistic assumption. However, PLSA is not a well-defined generative model for documents, because the model has no natural method to assign probability to a new document. LDA is an extension of PLSA. Assuming Dirichlet distribution weights for each unigram model, LDA can assign probability to new documents.

The DM is another MU extension. For DM, the assumption of MU — one topic for one document — remains, but distribution of unigram probability is directly modeled by the Dirichlet distribution. Other models, including PLSA and LDA, can reveal multitopic structure, but each topic model is quite simple — a unigram model. Though DM assumes a unitopic structure for each document, its topic models are richer than those of its strong rivals.

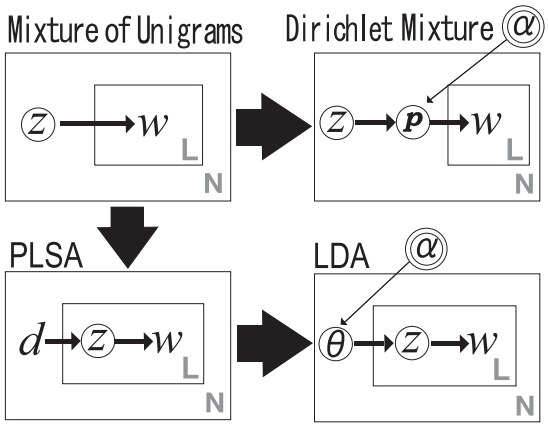

Figure 1: Graphical model expressions of generative text models

5 Experiments

The performance of the DM was compared with those of the LDA and the MU in test-set perplexities using adaptive $n$gram language models.

Training in all three models relied on the same training data and were evaluated with the same test data. The training data was a set of 98,211 Japanese newspaper articles from the year 1999. The test data was a set of 495 randomly selected articles of more than 40 words, from the year 1998. The vocabulary comprised the 20,000 most frequent words in the training data, and had a cover rate of 97.1%.

The variational EM method was used to estimate the LDA parameter (Blei et al. 2003), but $\boldsymbol{\alpha}$ in the Dirichlet parameter for LDA was updated with Minka's fixed-point iteration method (Minka 2003), instead of the Newton-Raphson method. For the DM, the estimation method (1) based on an LOO likelihood was used. Training for both models used the same stopping criteria — the change of closed perplexity for the training data before and after one global loop of iteration is less than 0.1%. Models were constructed with both LDA and DM having 1, 2, 5, 10, 20, 50, 100, 200, and 500 components.

Figure 2 presents the perplexity of each model for its different number of components. Each perplexity is computed as an inverse of probability per word, a geometrical mean of a document probability, that is, an evidence probability for data. For the DM, the probability of the Polya mixture for a document is the document probability. The DM consistently outperforms both other models. However, DM performance is best at 20 components, as it saturates with somewhat fewer components. DM is a generalized version of MU and the saturation suggests that DM suffers the MU overfitting problem.

Figure 3 presents the perplexity of the adaptive language models for different numbers of components. Adaptive language models predict the probability of the next word by using a history, such as a conventional $n$gram language model. They employ a longer history, such as an entire previously processed section, rather than just the two words preceding the target

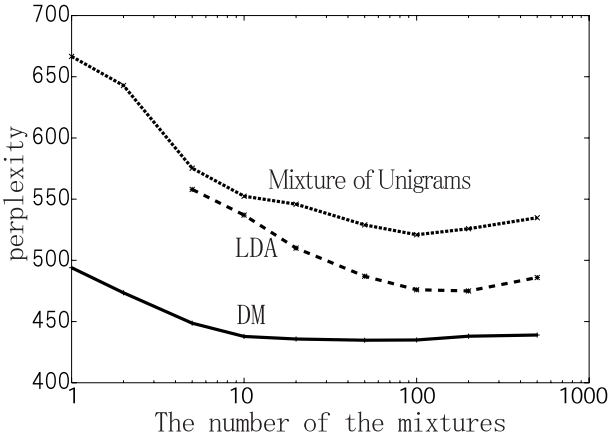

Figure 2: Comparison of test-set perplexity by document probability

word. In this experiment, the models were adapted to a section of a document from the first to the current word, and then probabilities were computed for the next 20 words. Every 20 words, this operation was repeated. Figure 3 also shows that DM outperforms better than LDA.

For the next experiment, a trigram language model was developed, dynamically adapted to longer histories for topic-based models. A unigram rescaling method was used for the adaptation (Gildea and Hofmann 1999). Figure 4 shows the perplexity of combined adaptive language models. Like the above experiments, the performance of a unigram-rescaling trigram model with DM is better than that with LDA.

Table 1 shows the best perplexities for each model. The values in parentheses are perplexity reduction rates from baseline models.

LDA is a multitopic text model, which reveals a mixture of topics in a document, while DM is a unitopic text model, which assumes one topic per document. Generally speaking, there are few documents with just one topic. This raises the question as to why DM is better than LDA in perplexity for those experiments. While there is no clear answer, it is possible that the training and test materials for those experiments were based on newspaper articles, and thus somewhat focused on a single topic. DM may capture the detailed distribution of word probabilities on topics using multiple Dirichlets, whereas LDA simply captures the distribution indirectly as a topic proportion or weight for a mixture using a single Dirichlet. The same experiments need to be conducted with web data, which contains more complex topical structures.

6 Conclusions

Finite mixture of Dirichlet distributions was investigated as a new text model. Parameter estimation methods were introduced, based on Minka's fixed-point methods (Minka 2003)

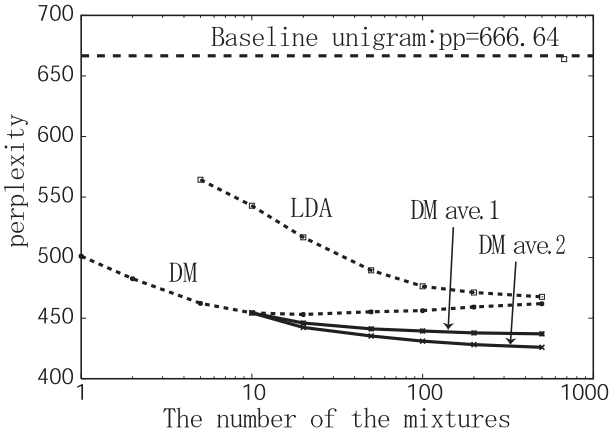

Figure 3: Comparison of test-set perplexity by history adaptation probability(unigram)

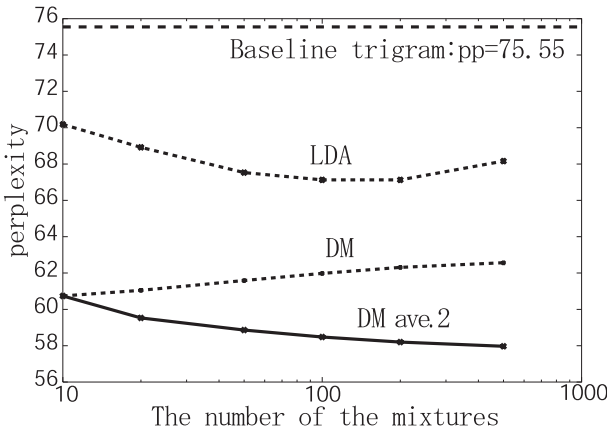

Figure 4: Comparison of test-set perplexity by history adaptation probability(trigram)

and the EM algorithm using the mixture of the Dirichlet-multinomial distribution. Experimental results for applications of statistical language models were demonstrated and their performance compared for perplexity. The DM model achieved the lowest perplexity level despite its unitopic nature.

References

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan. 2001. "Latent Dirichlet Allocation," Neural Information Processing Systems, vol.14.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol.3, pages 993-1022.

Table 1: Minimum perplexity of each aspect model

	history adaptation (unigram)	history adaptation (trigram)	document probability
DM	453.06(32.0%)	60.74(19.6%)	434.73
DM ave.2	425.97(36.1%)	57.97 (23.2%)	-
LDA	467.61(29.9%)	67.13 (11.1%)	474.82
Mixture of Unigrams	–	–	520.95

[3] Kenneth W. Church and William A. Gale. 1995. Poisson mixtures. Natural Language Engineering, Vol.1, No.1, pages 163–190.

[4] Kenneth W. Church. 2001. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In Proc. of Coling 2000, pages 180–186.

[5] A. P. Dempster, N. M. Laird and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. of Royal Statistical Society, Series B, Vol.39, pages 1–38.

[6] T. Hofmann. 1999. "Probabilistic latent semantic indexing," Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, pp.50-57, Berkeley, California.

[7] K. Sjölander, K. Karplus, M. Brown, R. Hunghey, A. Krogh, I.S. Mian, and D. Haussler. 1996. "Dirichlet mixtures:a method for improved detection of weak but significant protein sequence homology," Computer Applications in the Biosciences, vol.12, no.4, pp.327–345.

[8] D. Gildea, and T. Hofmann. 1999. "Topic-based language models using em," Proc. of the 6th European Conference on Speech Communication and Technology (EUROSPEECH'99).

[9] R.M. Iyer, and M. Ostendorf. 1999. "Modeling long distance dependence in language:topic mixtures versus dynamic cache models," IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, vol.7, no.1, pp.30–39.

[10] S.T. K. Nigam, A. McCallum, and T. Mitchell. 2000. "Text classification from labeled and unlabeled documents using EM," Machine Learning, vol.39, no.2/3, pp.103–134.

[11] T. Minka. 2003. "Estimating a Dirichlet distribution," http://www.stat.cmu.edu/ minka/papers/dirichlet/

[12] Ronald Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. Computer Speech and Language, Vol.10, No.3, pages 187–228.

A Derivation of the update formula of $\boldsymbol{\alpha}$ for the MLE

Minka's fixed-point iteration method (Minka 1999) and the EM algorithm were used. The following equations (Minka 2003) were used to get the lower bound of the second term of the equation (4):

$$\frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + y_i)} \geq \frac{\Gamma(\bar{\alpha}_m)\exp\left\{(\bar{\alpha}_m - \alpha_m)b_{im}\right\}}{\Gamma(\bar{\alpha}_m + y_i)},$$

and

$$\frac{\Gamma(\bar{\alpha}_{mv} + y_{iv})}{\Gamma(\bar{\alpha}_{mv})} \geq c_{imv}\bar{\alpha}_{mv}^{a_{imv}} \quad (if \quad y_{iv} \geq 1),$$

where

$$y_i = \sum_v y_{iv}, \qquad \alpha_m = \sum_v \alpha_{mv},$$

and

$$\bar{\alpha}_m = \sum_v \bar{\alpha}_{mv}.$$

The lower bound is:

$$\begin{aligned}
&\sum_i \sum_m P_{im} \log \frac{\Gamma(\bar{\alpha}_m)}{\Gamma(\bar{\alpha}_m + y_i)} \prod_{v=1}^{V} \frac{\Gamma(y_{iv} + \bar{\alpha}_{mv})}{\Gamma(\bar{\alpha}_{mv})} \\
&\geq \sum_i \sum_m P_{im} \left[\log \frac{\Gamma(\bar{\alpha}_m)\exp\left\{(\bar{\alpha}_m - \alpha_m)b_{im}\right\}}{\Gamma(y_i + \bar{\alpha}_m)} + \sum_v \log c_{imv}\alpha_{mv}^{a_{imv}} \right] \equiv Q'(\boldsymbol{\alpha}),
\end{aligned}$$

where

$$\begin{aligned}
a_{imv} &= \left\{\Psi(\bar{\alpha}_{mv} + y_{iv}) - \Psi(\bar{\alpha}_{mv})\right\}\bar{\alpha}_{mv}, \\
b_{im} &= \Psi(\bar{\alpha}_m + y_i) - \Psi(\bar{\alpha}_m), \\
c_{imv} &= \frac{\Gamma(\bar{\alpha}_{mv} + y_{iv})}{\Gamma(\bar{\alpha}_{mv})}\bar{\alpha}_{mv}^{-a_{imv}}.
\end{aligned}$$

This lower bound can be maximized with the following update formula for α_{mv}.

$$\frac{\partial Q'(\boldsymbol{\alpha})}{\partial \alpha_{mv}} = -\sum_i P_{im} b_i + \frac{1}{\bar{\alpha}_{mv}} \sum_i P_{im} a_{imv} = 0$$

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_i P_{im}\{\Psi(y_{ik} + \bar{\alpha}_{mv}) - \Psi(\bar{\alpha}_{mv})\}}{\sum_i P_{im}\{\Psi(y_i + \bar{\alpha}_m) - \Psi(\bar{\alpha}_m)\}}$$

B Derivation for the update formula of $\boldsymbol{\alpha}$ for the maximum LOO likelihood

Given datum $\boldsymbol{y}_i^{-v}$ in which one word v is left out of a document, the predictive probability $P(v^*|\boldsymbol{y}_i^{-v})$ for the word v is the following from the equation (9):

$$p(v^*|\boldsymbol{y}_i^{-v}) \fallingdotseq \sum_m P_{im} \frac{\alpha_{mv} + y_{iv} - 1}{\alpha_m + y_i - 1},$$

where

$$P_{im}^{-v} \fallingdotseq P_{im}.$$

The LOO log-likelihood, $\mathcal{L}_{loo}$, is

$$\mathcal{L}_{loo}(\boldsymbol{y}|\boldsymbol{\alpha}) \fallingdotseq \sum_i \sum_v y_{iv} \log \sum_m P_{im} \frac{\alpha_{mv} + y_{iv} - 1}{\alpha_m + y_i - 1}.$$

Since P_{im} is very nearly to 1 or 0 for almost all cases, the preceding equation can be transformed:

$$\mathcal{L}_{loo}(\boldsymbol{y}|\boldsymbol{\alpha}) \fallingdotseq \sum_i \sum_v \sum_m y_{iv} P_{im} \log \left(\frac{\alpha_{mv} + y_{iv} - 1}{\alpha_m + y_i - 1} \right)$$

Using the preceding equation, likelihood can be maximized independently for α_{mv} of the m-th Dirichlet component. The lower bound of the LOO log-likelihood $\mathcal{L}_{loo}^m$ for the m-th component is:

$$\mathcal{L}_{loo}^m \geq \sum_i \big(\sum_v y_{iv} P_{im} q_{imv} \log \alpha_{mv} - y_i P_{im} a_{im} \alpha_m \big) + (const.),$$

where

$$q_{imv} = \frac{\bar{\alpha}_{mv}}{\bar{\alpha}_{mv} + y_{iv} - 1},$$
$$a_{im} = \frac{1}{\bar{\alpha}_m + y_i - 1}.$$

The following inequalities (Minka 2003) were used to get the above bound:

$$\log(n + x) \geq q \log x + (1 - q) \log n - q \log q - (1 - q) \log(1 - q)$$

$$\text{where } q = \tfrac{\hat{x}}{n + \hat{x}},$$

$$\log(x) \leq ax - 1 + \log \hat{x}$$

where $a = 1/\hat{x}$.

An update formula for fixed-point iteration is:

$$\alpha_{mv} = \bar{\alpha}_{mv} \frac{\sum_i P_{im}\{y_{iv}/(y_{iv} - 1 + \bar{\alpha}_{mv})\}}{\sum_i P_{im}\{y_i/(y_i - 1 + \bar{\alpha}_m)\}}.$$