# 複数モデルの統合による LDA トピックモデルの高精度化

中村明 [1)]　津田裕亮 [2)]　松本忠博 [2)]　池田尚志 [2)]　速水悟 [2)]

1）三洋電機㈱ ヒューマンエコロジー研究所
2）岐阜大学 工学部 応用情報学科

## 1. はじめに

単語間の局所的な依存関係をモデル化するN-gramモデルは，シンプルであるが故に汎用性が高く，様々な自然言語処理タスクに応用されている．N-gramモデルに単語間の大域的な依存関係を組み込むことによって，言語をより精緻にモデル化することが可能であり，キャッシュモデルやトリガーモデルの併用はその一例である[1].

キャッシュモデルやトリガーモデルが単語間の長距離の依存関係を単語（対）の形で直接，モデル化するのに対し，近年，単語間の大域的な依存関係を話題（トピック）としてモデル化するトピックモデルの研究が進展している．unigramの混合モデルであるMixture of Unigrams[2]，潜在トピックを導入しLSI(Latent Semantic Indexing)を確率モデルとして再定式化したPLSI(Probabilistic Latent Semantic Indexing)[3]，PLSIのベイズ的発展形であるLDA(Latent Dirichlet Allocation)[4]，単語生起確率を混合ディリクレ分布に従う確率変数とするDM(Dirichlet Mixture)[5]などが提案されている．

これらトピックモデルでは文脈に基づいてunigram確率を動的に推定することによりパープレキシティを削減できる．そしてunigram rescaling[6]等の補間手法によってN-gramモデルと組み合わせることが可能であり，連続音声認識，同音異義語の誤り検出などへの適用が試みられている[7-9]．またBleiらはLDAを文書分類のための特徴抽出器としても用いている[4]．筆者らの一部は以前，N-gramモデルとキャッシュモデルを用いたテキスト入力支援システムを構築し，医療文書の入力支援における有効性を確認しているが[10]，トピックモデルを用いることによってこれを更に高精度化することを目指している．

言語モデルに限らず，有限個の事例に基づいてパラメータを推定する学習器では，一般にモデルのパラメータ数増加に伴ってopen dataに対する精度（汎化能力）が低下する，いわゆる過適応が不可避である．N-gramにおけるスパースネスの問題もこの一種であるし，PLSIやLDA等のトピックモデルにおいても潜在トピック数の増加に伴い過適応が問題となる．そのため，計算コストを考慮しつつ最適なトピック数を決定する必要がある．

トピックモデルにおいて過適応の問題に対処した先行研究としては，DM(Dirichlet Mixture)のパラメータ推定において階層ベイズモデルを用いて平滑化を行った報告がある[11]．文献[5]では，モデル規模（混合ディレクレ分布の混合数）の異なる複数のDMを重み付きで平均することにより過適応を抑制できることが示されている．また，階層ディリクレ過程を用いてLDAの最適な潜在トピック数を決定する手法も提案されている[12].

これに対し本稿では，独立に学習した複数のLDAを組み合わせることによって，単一のLDAよりも高い推定精度を実現することを目的とする．複数のモデルを組み合わせる点で上述の文献[5]に述べられた「モデル平均」の方式と類似するが，本稿で議論する方式はベースとなるモデルがLDAである以外に(1)比較的小規模かつ同一規模の複数のモデルを統合する，(2)リサンプリングにより生成した異なるサブセットを学習文書に用いた場合の効果を検証する，(3)同等の計算コスト（各モデルの潜在トピック数の合計）の元での性能比較を行う，といった点が異なる．

複数の学習器を組み合わせて精度を向上する枠組みはensemble learning，voting network，committee等様々な名称で呼ばれるが，その本質は統合によるモデルの拡大が汎化能力向上をもたらす点にあり，MoE(Mixture of Experts)[13]，bagging[14]，boosting[15]など様々なアルゴリズムが考案されている．

本稿で提案するシステムは，トピックモデルに複数学習器統合の考え方を取り入れたものであり，独立に学習した複数個のLDAモデルによる推定結果を統合することにより，最終的なunigram確率を出力する．これによって，モデル規模（潜在トピック数×モデル数）を大きくしていっても性能が低下しないこと，全体のモデル規模が同程度の単一モデルよりも常にパープレキシティを削減できること等を実験的に示す．以下，2章でLDAの概要について述べ，3章で提案システムの構成を説明する．そして4章で評価実験の結果を示し5章で考察を行う．

## 2. LDAの概要

LDA(Latent Dirichlet Allocation)[4]は，各潜在トピック($z_1, z_2,...,z_C$)($C$:潜在トピック数)の生成確率$\boldsymbol{\theta}=(\theta_1, \theta_2,..., \theta_C)$が多項分布の共役事前分布であるディリクレ分布$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$に従うと仮定したモデルである．文書$\boldsymbol{d}=(w_1, w_2,..., w_{|d|})$の出現確率は次式で表される（$|d|$は文書$\boldsymbol{d}$の総単語数を表す）．

$$p(d\,|\,\boldsymbol{\alpha},\boldsymbol{\beta}) = \int \mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\left(\prod_{n=1}^{|d|}\sum_{k=1}^{C} p(w_n\,|\,z_k,\boldsymbol{\beta})\,p(z_k\,|\,\boldsymbol{\theta})\right)d\boldsymbol{\theta} \quad (1)$$

$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$がLDAのモデルパラメータであり，$\beta_{kj}$はトピック$z_k$における語$w_j$のunigram確率$p(w_j|z_k)$を表す($1 \leqq j \leqq V$)( $V$:語彙数)．$\boldsymbol{\alpha}=(\alpha_1, \alpha_2,...,\alpha_C)$はディリクレ分布

$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{C}\alpha_k\right)}{\prod_{k=1}^{C}\Gamma(\alpha_k)}\prod_{k=1}^{C}\theta_k^{\alpha_k-1} \qquad (2)$$

のパラメータである．パラメータ$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$の学習には変分ベイズ法による近似計算が用いられる[4].

未知の文脈$h$に対する文脈適応は，学習時と同様の変分近似により計算する．即ち，$h$に対する変分パラメータ$\gamma_k$および$\phi_{kj}$を導入し，学習済みの$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$を用いて以下の手順を収束

するまで繰り返す.

$$\mathrm{VB-Estep}: \quad \phi_{kj} \propto \beta_{kj} \exp(\Psi(\gamma_k) - \Psi(\sum_{k'=1}^{C} \gamma_{k'})) \qquad (3)$$

$$\mathrm{VB-Mstep}: \quad \gamma_k = \alpha_k + \sum_{j=1}^{V} n(h, w_j)\phi_{kj} \qquad (4)$$

$\Psi(\gamma)$はdigamma関数であり，$n(h,w_j)$は$h$における語$w_j$の出現回数を表す．得られた$\gamma_k$を文脈$h$の元での各潜在トピックの混合比とする．したがって，文脈$h$の元での語$w_{j'}$の生起確率は次式により与えられる．

$$p(w_{j'} \mid h) = \frac{\sum_{k=1}^{C} \gamma_k \beta_{kj'}}{\sum_{k=1}^{C} \gamma_k} \qquad (5)$$

LDAはトピックの事前分布にディリクレ分布を用いることにより，トピックの拡がりやトピック間の関係を表現できる点でPLSIより優れている．またベイズ推定に基づくため過適応の問題が少ないとされている．

## 3. 提案システム

本稿で提案するシステムの構成を図1に示す．本システムは独立に学習した$M$(≧2)個のLDAモデル($Q_1, Q_2,\cdots,Q_M$)を持つ．各モデルの潜在トピック数は必ずしも同じでなくてもよいが，本稿では各モデルの構成を同一とし学習条件の違いによる統合の効果を検証するため，各モデルの潜在トピック数が同じ場合のみを扱う．従って本稿では各モデル間の差異は，後述のように学習サンプルの違いと学習時の初期値の違いのみによる．以降，複数個のLDAからなる図1の構成を**m-LDA**(multiple-LDA)と呼ぶ．

$L$形態素からなる文脈$h$が入力されると，各LDAモデル($Q_1, Q_2,\cdots,Q_M$)はそれぞれ文脈推定を行い，(5)式により$h$の元での語$w$の生起確率$p_m(w|h)$を求める($1 \leqq m \leqq M$)．$M$個の$p_m(w|h)$の平均値

$$\bar{p}(w \mid h) = \frac{1}{M} \sum_{m=1}^{M} p_m(w \mid h) \qquad (6)$$

を最終的な語$w$の生起確率の推定値として出力する．$p_m(w|h)$を適当に重み付けして平均する方式も考えられるが本稿では扱わない．

## 4. 評価実験

### 4.1. 学習データおよび評価データ

学習用データおよび評価用データを以下に示す．

［学習用データ］
CD－毎日新聞2005データ集[16]全記事の約1/2（48035件;記事番号下1桁が奇数のもの）．のべ約1439万形態素，異なり語数141666

［評価用データ］
CD－毎日新聞2006データ集[16]の内，200文字以上の記事から無作為抽出した1000件．のべ約40万形態素

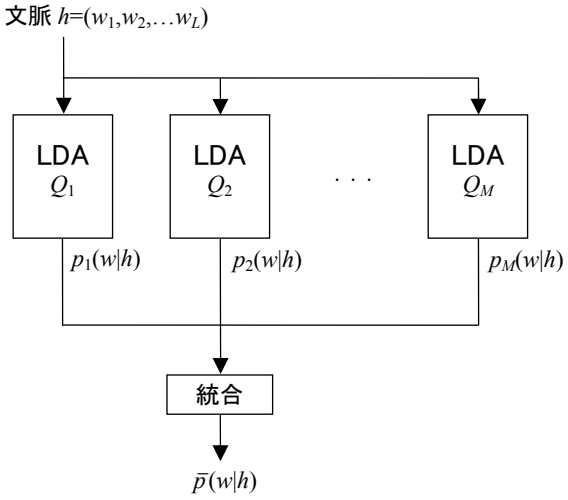



**図1．システム構成**

学習用データ・評価用データとも，文節構造解析システムibukiC[17]により形態素解析を行った．評価データ中，学習データに含まれない未知語はのべ2411語であった．

### 4.2. 学習条件

潜在トピック数$C$=20/30/50/70の場合に対し，それぞれ以下に示す2通りの方法でLDAの学習を行った．

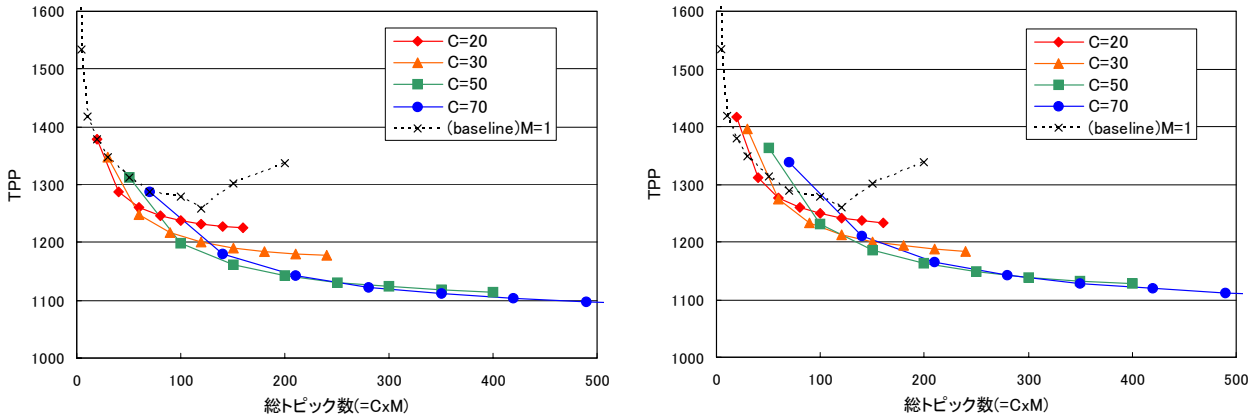
(1)**m-LDA1**（同一の学習データセット使用）：前節で示した学習用データ48035件（以降，学習データセット$\boldsymbol{D}_0$と呼ぶ）を学習データセットとし，$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$に異なる初期値を与えて$M$回の学習を行い，$M$個のLDAを構築する．

(2)**m-LDA2**（異なる学習データセット使用）：$\boldsymbol{D}_0$から復元抽出により48035サンプルを抽出する作業を$M$回行って$M$個のデータセット($\boldsymbol{D}_{b1}, \boldsymbol{D}_{b2},\ldots,\boldsymbol{D}_{bM}$)を作成，これらを学習データセットとして$M$回の学習を行い，$M$個のLDAを構築する．

なお(2)で用いる$M$個のデータセット($\boldsymbol{D}_{b1}, \boldsymbol{D}_{b2},\ldots,\boldsymbol{D}_{bM}$)の組は，すべての潜在トピック数に対して共通のものを用いた．上記(1)(2)とも，元の学習データセット$\boldsymbol{D}_0$における出現回数が2回以下の語を除いた75314語で学習を行った．学習アルゴリズムは2章で述べた変分ベイズ法を使用し，パラメータ$\boldsymbol{\alpha}$の推定にはFixed-point iteration[18]を用いた．収束判定は，学習データセットに対する1ステップ前からのパープレキシティの減少幅が0.5未満となった時点で収束とした．

### 4.3. 評価方法

評価用データセットに対して提案システムによりunigram確率を求め，テストセットパープレキシティ（*TPP*）で評価する．LDAに与える文脈$h$の長さ$L$は予備実験により20形態素とした．評価用データセット中の各記事の先頭20形態素まではトピック推定を行わず，文脈非依存のunigram確率をそのまま用いる．




(a)m-LDA1　　(b)m-LDA2

図2. モデル数・トピック数の違いによる TPP の推移

### 4.4. 実験結果

各LDAの潜在トピック数$C$=20/30/50/70の場合についてモデル数$M$=8までの評価を行った結果を図2に示す．システム全体でのモデル規模が同等の条件下で比較するため，横軸は総トピック数（=潜在トピック数$C$×モデル数$M$）とした．$C$=20/30/50/70の各プロットは，8個のモデルから$M$個（$1 \leqq M \leqq 8$）を選ぶすべての組み合わせ（${}_8C_M$通り）についての平均値である．baselineはデータセット$\boldsymbol{D}_0$で学習した単一のLDA（即ち$M$=1）で$C$を変化させた場合の性能を示す．なお，図示されていないが文脈非依存のunigramによるテストセットパープレキシティ（$TPP_0$）は1905であった．

図2に示すように，**m-LDA1**，**m-LDA2**いずれの場合もモデル数の増加に伴い$TPP$が単調に減少する．baselineでは過適応のため$C$=120を境に性能が悪化するのに対し，提案システムでは過適応による性能低下が抑えられ，単一のLDAに比べ大幅に高い推定精度が得られている．$M \geqq 2$ではすべての場合においてbaselineの性能を上回っており，複数LDAの統合によって同規模の（即ち$C \times M$が等しい）単一LDAより常に性能が改善することが分かる．**m-LDA1**と**m-LDA2**とを比較すると，**m-LDA2**のほうが統合による$TPP$の減少がやや大きい．ただし統合前（$M$=1）の性能が**m-LDA1**に比べやや劣るため，統合後の性能はほぼ同程度である．

次に，システム構成の違いによる評価用データセット中の各記事に対する$TPP$の傾向を図3と図4に示す．図3の縦軸は，それぞれの構成における各記事の$TPP$を$C$=50，$M$=1の場合を基準とした相対値で表しており，下にいくほど$C$=50，$M$=1の場合より推定精度が改善されていることを意味する．図4の縦軸は同様に$C$=20，$M$=1の場合を基準とした相対値である．横軸は文脈非依存のunigramによる各記事に対する$TPP_0$であり，$\overline{TPP}$ は全記事に対する$TPP$を示す．

図3上段は**m-LDA1**で$C$=50に固定し$M$を2,3,4とした場合，下段は$M$=1とし（単一LDA）$C$=100,150,200と変化させた場合を表す．図より，**m-LDA1**ではモデル数の増加に伴ってほとんどの記事に対して推定精度が向上しており，複数モデル統合によって性能が安定化することが分かる．一方，単一モデルのままトピック数を増やすと縦軸方向のバラツキが大きくなり，記事によって当たり外れの大きい不安定な特性を示すようになる．特に$TPP_0$が大きい（もともと予測が難しい）記事で悪化する傾向がある．$C$=100，$M$=1では$C$=50，$M$=1に比べ全記事に対する$TPP$は低いにも関わらず（図2(a)参照），記事ごとに見ると$C$=50，$M$=1の場合より精度が低下するケースがかなりあることも分かる．

図4は$\overline{TPP}$がほぼ等しい構成での記事ごとの$TPP$の違いを示す．$C$=20，$M$=3と$C$=120，$M$=1とでは全記事に対する平均推定精度は同等であるにも関わらず，$C$=120，$M$=1の場合は記事によって精度のバラツキが大きく，$C$=20，$M$=1の場合より悪化するケースも少なくない．対して$C$=20，$M$=3では安定して精度が向上している．

## 5. 考察

実験の結果，複数のトピックモデルから得られる推定結果を統合することによって過適応を抑制し性能を向上・安定化できることが分かった．複数の学習器を統合する場合，一般には個々の学習器が異なった傾向を持つほど効果が大きいとされている．**m-LDA2**が**m-LDA1**に比べ統合による効果が大きいのは，別々の学習データを用いることによりモデル間の差異がより大きくなったためである．しかし今回の実験では**m-LDA1**と比べ効果の差はわずかであった．このことから，LDAでは初期値の違いから異なる局所解に収束することによるモデル間の差異のほうが，学習データの違いに起因する差異よりもはるかに大きいと考えられる．言い換えれば，初期値の違いによって異なる学習結果に収束する不安定さをモデル統合によって軽減できていると言える．

各記事に対する$TPP$を比較した結果からは，複数モデル統合により記事ごとの推定精度が安定して向上することが確かめられた．単一モデルで潜在トピック数を増やしていくと文脈によって当たり外れが大きくなるが，小規模なモデルを統合することによりこれを抑制することができる．

今回実験した範囲では，各LDAの潜在トピック数$C$が大きいほど統合時の精度は高いが，さらに$C$が大きい場合については追加実験が必要である．$C$が大きくなるにつれて差は縮

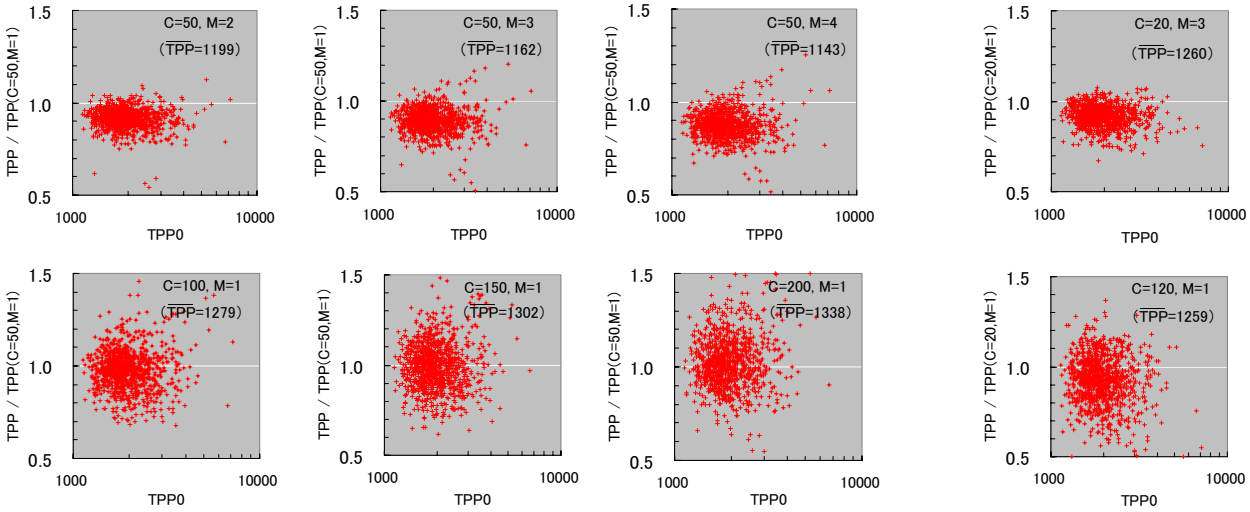



図3. 記事ごとの推定精度（1）

（モデル数・トピック数の一方のみを変化させた場合； 上段：C=50, 下段：M=1）

図4. 記事ごとの推定精度（2）

（同程度の$\overline{\text{TPP}}$での比較）

小しているため，統合の際の最適な潜在トピック数を見極める必要がある．

計算コストについては定量的な比較を行っていないが，適応時の各LDAでの演算量は潜在トピック数とともにほぼリニアに増加するため，$C \times M$が同じ場合により高い精度が得られる提案システムが有利である．学習時の計算時間は潜在トピック数が増えると大きく増える傾向があるため，システム全体の規模が同等の場合，複数モデルを用いるほうが学習に要する計算時間は短く済む．

## 6. まとめ

トピックモデルの推定精度向上を目的として，複数のLDAからなる構成（**m-LDA**）を提案，評価を行った．その結果，提案システムでは過適応による性能低下が抑制され，同規模の単一LDAより常に性能が改善すること，モデル規模が大きくなっても性能が向上し続けること，統合により推定精度が安定化することが確かめられた．また，異なる学習データで学習したモデルを統合した場合により効果が得られることが分かった．

今後は，複数モデルを統合する際の最適な潜在トピック数や学習条件に関し検討するとともに，unigram rescaling等を用いたN-gram推定精度の評価，テキスト入力支援などのアプリケーションに適用した場合の評価を行う予定である．さらに，同様の構成でPLSI，Dirichlet mixtureなど他のトピックモデルを用いた場合についても比較・検証を行う．

## 参考文献


[1] 北研二, “確率的言語モデル”, 東京大学出版会, 1999.

[2] S. Thrun, K. Nigam, A. McCallum and T. Mitchell, “Text classification from labeled and unlabeled documents using EM”, Machine Learning, Vol.39, No.2/3, pp.103–134, 2000.

[3] T. Hofmann, “Probabilistic latent semantic indexing”, Proc. of 22nd Annual ACM Conference on Research and Development in Information Retrieval, pp.50–57, 1999.

[4] D. Blei, A. Y. Ng and M. Jordan, “Latent dirichlet allocation”, Neural Information Processing Systems, Vol.14, 2001.

[5] 貞光九月，三品拓也，山本幹雄, “混合ディリクレ分布を用いたトピックに基づく言語モデル”, 電子情報通信学会論文誌D-II Vol.J88-D-II, No.9, pp.1771-1779, 2005.

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

[7] 高橋力矢，峯松信明，広瀬啓吉，“複数のバックオフN-gramを動的補間する言語モデルの高精度化”，情報処理学会研究報告SLP-49-11, pp.61-66, 2003.

[8] 根本雄介，秋田祐哉，河原達也, “講義音声認識のためのスライド情報を用いた言語モデル適応”，言語処理学会第13回年次大会論文集, pp.131-134, 2007.

[9] 三品拓也，貞光九月，山本幹雄, “確率的LSAを用いた日本語同音異義語誤りの検出・訂正”，情報処理学会論文誌 Vol.45, No.9, pp.2168-2176, 2004.

[10] 中村明，川尻博光，金川誠，松本忠博，池田尚志，速水悟，紀ノ定保臣，“統計的言語モデルに基づく電子カルテ入力支援システムの開発”，言語処理学会第13回年次大会論文集, pp.998-1001, 2007.

[11] 貞光九月，待鳥裕介，山本幹雄, “混合ディリクレ分布パラメータの階層ベイズモデルを用いたスムージング法”，情報処理学会研究報告 SLP53-1, pp.1-6, 2004.

[12] Y. Teh, M. Jordan, M. Beal and D. Blei, “Sharing clusters among related groups: Hierarchical dirichlet process”, NIPS 2004, MIT Press, 2004.

[13] R. Jacobs, M. Jordan, S. Nowlan and G. Hinton, “Adaptive mixtures of local experts”, Neural Computation, Vol.3, pp.79-87, 1991.

[14] L. Breiman, “Bagging predictors”, Technical Report 421, Statistics Department, Univ. of California, Berkeley, 1994.

[15] R. Schapire, “The strength of weak learnability”, Machine Learning, Vol.5, pp.197-227, 1990.

[16] CD-毎日新聞2005/2006データ集

[17] 山田佳裕，脇田貴之，大口智也，池田尚志，“文節構造解析システムibukiCの解析仕様および精度の比較と評価”，言語処理学会第13回年次大会論文集, pp.167-170, 2007.

[18] T.Minka, “Estimating a Dirichlet distribution”, http://www.stat.cmu.edu/~minka/papers/dirichlet/, 2003.