LDA の文脈長最適化によるトピック依存 N-gram の高精度化

中村明 [1)]　速水悟 [2)]

1）三洋電機㈱ エコロジー技術研究所　　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近年，進展しており，PLSI(Probabilistic Latent Semantic Indexing)[1]，LDA(Latent Dirichlet Allocation)[2]，DM(Dirichlet Mixture)[3]などのモデルが提案されている．トピックモデルでは話題に基づいて unigram 確率を適応的に推定できる．そして unigram rescaling[4]等の補間手法によって *N*-gram を高精度化することが可能であり，連続音声認識，同音異義語のかな漢字変換誤り検出，予測入力によるテキスト入力支援などへ適用されている[5-7]．

テキスト入力支援や連続音声認識等のオンラインアプリケーションにトピックモデルを適用する場合，入力テキストの話題変化に逐次，適応すること（以下，これをオンライン適応と呼ぶ）が必要である．オンライン適応では文脈長，すなわち適応時に文脈として用いる単語列（ヒストリ）の長さが精度に影響するが，先行研究の多くでは文書先頭以降の全単語を用いる[3,4]か，文脈長を一定単語数とする方法[5,7]が採られてきた．しかし実際のアプリケーションでは文書の境界が明示的に与えられるとは限らず，文書内でもトピックが刻々と変化する上，その速さも一定ではない．したがって入力に応じて適切な文脈長を動的に選択できることが望ましい．

文献[8]では Particle Filter を用いてトピックの変化点を確率的に推定することにより様々な文脈長からの予測を混合する方式が提案されている．この方式ではDMにおいてパープレキシティを 6～10%程度削減しているが，LDA では精度向上はわずかであった．一方，文献[9]では LDA においてトピック混合比に基づいて隣接する文書ブロック間の類似度を評価してトピック変化点を推定することによりパープレキシティを削減できることが示されている．しかし，トピック変化点を検出することがオンライン適応を高精度化する最良の方法であるかどうかは必ずしも自明ではない．

これに対し本稿では，現在の文を最も精度よく推定できるヒストリ始点を逐次求め，このヒストリ始点以降を用いて次の語を予測する方式を提案する．本方式では現在の予測対象語を推定する上で最適な文脈長をより直接的に選択でき，独立に学習した複数のLDAによる推定結果を重み付きで統合することによりさらに精度を向上できる．これにより，精度と計算コストの面で既存の方式と比較して同等以上の性能が得られることを示す．

2. LDA(Latent Dirichlet Allocation)

2.1. LDAの概要

LDA[2]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$\boldsymbol{\theta}=(\theta_1, \theta_2,..., \theta_C)$がディリクレ分布Dir($\boldsymbol{\theta}|\boldsymbol{\alpha}$)に従うと仮定したモデルである．文書$d=(w_1, w_2,..., w_{|d|})$の出現確率は次式で表される（$|d|$は文書$d$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，β_{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Dir($\boldsymbol{\theta}|\boldsymbol{\alpha}$)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[2]やMCMC(Markov Chain Monte Carlo)[10]が用いられるが，本稿では変分ベイズ法を用いる．

2.2. LDAのオンライン適応

未知の文脈$\boldsymbol{h}$に対するトピック適応は，学習時と同様の変分近似により計算する．即ち，変分パラメータγ_kおよびϕ_{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．

$$\mathrm{VB-Estep}: \quad \phi_{kj} \propto \beta_{kj} \exp(\Psi(\gamma_k)) \quad (2)$$

$$\mathrm{VB-Mstep}: \quad \gamma_k = \alpha_k + \sum_{j=1}^{V} n(\boldsymbol{h}, w_j)\phi_{kj} \quad (3)$$

$\Psi(\gamma)$はdigamma関数であり，$n(\boldsymbol{h},w_j)$は$\boldsymbol{h}$における語w_jの出現回数を表す．$\boldsymbol{h}$の元での語$w_{j'}$の生起確率は，γ_kをトピック混合比として次式により求められる．

$$p(w_{j'}\,|\,\boldsymbol{h}) = \frac{\sum_{k=1}^{C}\gamma_k\beta_{kj'}}{\sum_{k=1}^{C}\gamma_k} \quad (4)$$

$\boldsymbol{h}$を固定長とする場合，新聞記事コーパスを対象とした実験では1文程度，即ち約20形態素が望ましいと指摘されている[5]．しかし最適な文脈長は逐次変化すると考えられるため，入力に応じて文脈長を動的に制御できることが望ましい．

3. 提案方式

3.1. 文脈長の最適化

提案方式において文脈長を最適化する手順を図 1 を用いて説明する．

語w_tは現在時刻tにおける予測対象語であり，w_{s_0}はw_tを含む文の先頭の語である．w_{s_k}はk文前の文の先頭の語であり($1 \leqq k \leqq K$)，$s_K < \ldots < s_1 < s_0 < t$である．直前の予測対象語$w_{t-1}$までが既知であるとして，各文の先頭すなわち$\{w_{s_K}, \ldots, w_{s_1}, w_{s_0}\}$の中から現在の文脈に最適なヒストリ始点を選ぶことを考える．

語w_{s_k}をヒストリ$\boldsymbol{h}$の始点b_hとして，現在の文の先頭w_{s_0}から直前の語 w_{t-1} までを順次，予測した場合における単語列$w_{s_0}^{t-1} = w_{s_0}\cdots w_{t-1}$の生成確率は

$$p(w_{s_0}^{t-1}\,|\,b_h = w_{s_k}) = \prod_{\tau=s_0}^{t-1} p_L(w_\tau\,|\,\boldsymbol{h} = w_{s_k}^{\tau-1}) \quad (5)$$

となる．ここで$P_L(w|\boldsymbol{h})$は$\boldsymbol{h}$を入力ヒストリとしてLDAにより求めた語wのトピック依存unigram確率である．式(5)による生成確率が大きいほど，w_{s_k}を始点とする$\boldsymbol{h}$は現在の文のうち既知の部分すなわち$w_{s_0}^{t-1}$を精度よく推定できる．このとき現在の予測対象語w_tも精度よく予測できる可能性が高いと考え，$\{w_{s_K}, \ldots, w_{s_1}, w_{s_0}\}$の中で式(5)の生成確率が最大となる候補を現在時刻tにおける最適なヒストリ始点とする．すなわち，時

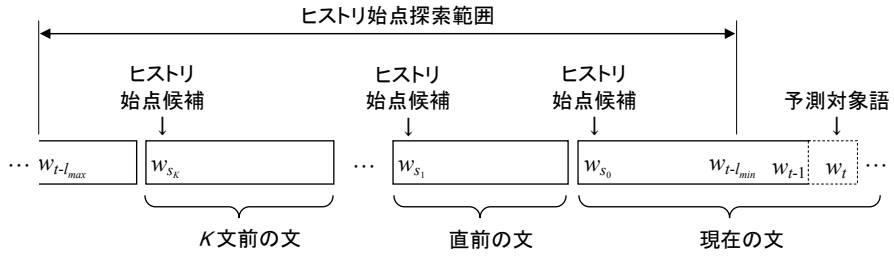

図 1 提案手法における文脈長の最適化

刻tにおける最適なヒストリ始点$\hat{b}_h(t)$を次式から求め，

$$\hat{b}_h(t) = \underset{b_h}{\operatorname{argmax}}\, p(w_{s_0}^{t-1} \mid b_h = w_{s_k}) \tag{6}$$

$\hat{b}_h(t)$をヒストリ始点として求めた$p_L(w_t \mid \boldsymbol{h} = \hat{b}_h(t) \cdots w_{t-1})$を予測対象語$w_t$のトピック依存 unigram 確率とする．

式(5)の生成確率の計算において，単語列$w_{s_0}^{t-1}$中の各語のトピック依存unigram確率を求めるには$\boldsymbol{h} = w_{s_k}^{\tau-1}$をヒストリとして$w_{s_0}^{t-1}$の語数分（すなわち$(t-s_0)$回）トピック適応を行う必要がある．この計算は式(2)，(3)の変分ベイズ法を含むため計算コストが大きい．しかしw_tが直前の語w_{t-1}と同じ文に属していれば（つまりw_tが文頭でない限り），同じヒストリ始点w_{s_k}に対して時刻$(t-1)$で計算した$w_{s_0}^{t-2}$の生成確率$p(w_{s_0}^{t-2} \mid b_h = w_{s_k})$を保持しておくことにより式(5)の計算は効率よく実行できる．すなわち

$$p(w_{s_0}^{t-1} \mid b_h = w_{s_k}) = p_L(w_{t-1} \mid \boldsymbol{h} = w_{s_k}^{t-2}) \cdot p(w_{s_0}^{t-2} \mid b_h = w_{s_k}) \tag{7}$$

として最後の語 w_{t-1} に対してのみトピック適応とトピック依存 unigram 確率の計算を行えばよい．したがって各時刻においてトピック適応を行う回数はヒストリ始点ごとに 1 回で済む．

なお，ここでは各文の先頭のみをヒストリ始点の候補としているが，原理的には現在の文の先頭w_{s_0}より過去の任意の位置をヒストリ始点とすることも可能である．また実装上は図1に示すようにヒストリ始点候補の探索範囲を現在時刻tより前のl_{min}～l_{max}形態素に限定する（$l_{min} < l_{max}$）．

以上のように，トピック変化点を検出してこれをヒストリ始点とする従来の方式がやや間接的であるのに対して，本手法ではより直接的な方法で現在の予測対象語を推定する上で最適な文脈長を求める．

3.2. 複数LDAの統合

文献[7]では独立に学習した複数の LDA による推定結果を統合することにより精度を向上・安定化できることが示されている．本稿では前節で示した文脈長の最適化を複数の LDA で別々に行い，これらの結果を統合することにより高精度化を図る．

まず前節の方法によりM個の LDA でそれぞれ最適なヒストリ始点$\hat{b}_h(t;m)$を求める（$1 \leqq m \leqq M$）．そして式(5)による既知単語列の生成確率$p^{(m)}(w_{s_0}^{t-1} \mid b_h = \hat{b}_h(t;m))$から得られる次式を各モデルによるヒストリ始点の尤度とする．

$$L(t,m) = \log \frac{p^{(m)}(w_{s_0}^{t-1} \mid b_h = \hat{b}_h(t;m))}{\prod_{m'=1}^{M} p^{(m')}(w_{s_0}^{t-1} \mid b_h = \hat{b}_h(t;m'))} \tag{8}$$

右肩の(m)はm番目の LDA から求めた確率を表し，既知単語列の生成確率が大きいモデルほど尤度も大きくなる．

次に各モデルで$\hat{b}_h(t;m)$をヒストリ始点としてw_tのトピック依存 unigram 確率$p_L^{(m)}(w_t \mid \boldsymbol{h} = \hat{b}_h(t;m) \cdots w_{t-1})$を求め，そしてこれを式(8)の尤度で重みづけし統合した結果を複数 LDA によるw_tのトピック依存 unigram 確率とする．

$$p_L(w_t \mid \boldsymbol{h}) = \sum_{m=1}^{M} L(t,m)\, p_L^{(m)}(w_t \mid \boldsymbol{h} = \hat{b}_h(t;m) \cdots w_{t-1}) \Big/ \sum_{m=1}^{M} L(t,m) \tag{9}$$

3.3. *N*-gram のトピック適応

LDAの学習とは別に，学習テキストからあらかじめ（トピック非依存の）N-gramを構築しておく．そして前節に示したトピック依存 unigram 確率$p_L(w_t|\boldsymbol{h})$を用いて，次式の unigram rescaling[4] によりトピック依存 N-gram 確率を計算し N-gram($N \geqq 2$)をトピックに適応させる．

$$p(w_t \mid w_{t-N+1}^{t-1}, \boldsymbol{h}) \propto \frac{p_L(w_t \mid \boldsymbol{h})}{p(w_t)}\, p(w_t \mid w_{t-N+1}^{t-1}) \tag{10}$$

$p(w_t)$：トピック非依存unigram確率

$p(w_t \mid w_{t-N+1}^{t-1})$：トピック非依存$N-$gram確率

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[11] 全記事(95,881 件)．約 2864 万形態素，異なり語数 185,196

［評価用データ］

CD-毎日新聞 2006 データ集[11] のうち，200 文字以上の記事から無作為抽出した 1000 件

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[12]により形態素解析を行った．評価用データについては，上記のデータから後述のように話題変化速度の異なる3個のデータセットを生成した．

4.2. 学習条件

前節で示した学習用データを用いてLDAの学習を行った．潜在トピック数C=100とし，語彙は学習用データにおける出現回数が4回以下の語を除いた77794語とした．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，$\boldsymbol{\alpha}$の推定には Fixed-point iteration[13]を用いた．収束判定は，学習用デー

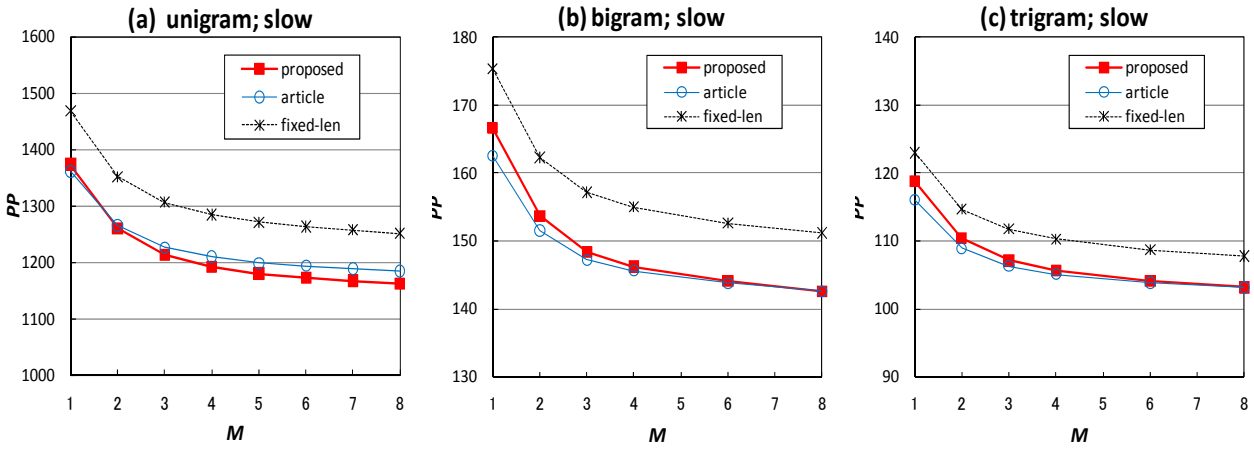

図2　提案手法におけるパープレキシティの変化

タに対する1ステップ前からのパープレキシティの減少幅が0.1%未満となった時点で収束とした．複数のLDAの統合による高精度化を評価するため，異なる初期値を与えて繰り返し学習を行い8個のLDAを構築した（すなわちモデル数Mの最大値を8とした）．

さらに同じ学習用データを用いてトピック非依存のN-gram(N≦3)を構築した．bi-gramとtri-gramの推定にはKneser-Neyスムージング[14]を用いた．

4.3. 評価方法

テキストの動的なトピック変化に対する提案手法の振る舞いを評価するためには，様々な速度で話題が変化していくテキストが必要となる．そこで先行研究[5,8]にならい，比較的長い元テキストから文をサンプリングすることにより話題変化速度の異なる3セットの評価テキストを作成した．元テキストは4.1節で示した評価用データ1000記事であり，文献[5]と同様，以下の手順で行う．

(1)元テキスト1000記事を掲載面コード[*1]に基づいて16個のサブテキストに分割し，各サブテキストの先頭を読み取り位置とする．
(2)一つのサブテキストをランダムに選び，現在の読み取り位置から連続するX文を採取する．
(3)Y文だけスキップして読み取り位置を進める．
(4)所定数の文が得られるまで(2)(3)を繰り返す．

ここでX, Yは表1に従う一様乱数であり，raw，slow，fastの順に話題変化が速くなる．採取する文の数は3セットとも2000文とした．

表1　評価テキストのパラメータ
Table 1　Properties of the test sets

セット名	パラメータ
raw	X=100, Y=0
slow	1≦X≦10, 1≦Y≦10
fast	1≦X≦3, 1≦Y≦10

[*1] 使用したコーパスの各記事には，社会・国際・スポーツ・文化など掲載面を示す16種の掲載面コードのうち一つが付与されている．

5. 実験結果

前章で示した評価テキストを用いてトピック依存N-gram（N≦3）のパープレキシティ（以下，PP）により提案手法の評価を行った．ヒストリ始点の探索範囲は予備実験によりl_{min}=5, l_{max}=100と定め，ヒストリ始点候補は各文の文頭に限定した．紙面の都合上，評価セットslowに対する結果のみ図2に示す．横軸は複数LDA統合時のモデル数Mを表しており，各モデルのトピック数Cは100トピックである．提案手法との比較のため，文脈長を20形態素に固定した場合（fixed-len）と，直近の記事境界をヒストリ始点とした場合（article）を併せて図示した．なお，文脈長を20形態素に固定した場合と直近の記事境界をヒストリ始点とした場合において，M≧2の時には個々のLDAによるトピック依存N-gram確率を重みなしで平均する方式[10]を用いた．

図より，提案手法では文脈長固定の場合よりもPPを削減できている．また統合モデル数Mの増加に伴い直近の記事境界をヒストリ始点とした場合と同程度までPPを削減できており，提案手法では記事境界が未知であっても記事境界が既知の場合と同等の推定精度が得られることが示されている．

図3は評価セットslowに対して，4個のLDA統合時（M=4）に提案手法によって得られた各単語のトピック依存unigram確率を，同じくM=4で文脈長を20形態素に固定した場合と比較した結果である．横軸は文脈長固定の場合のトピック依存unigram確率，縦軸は文脈長固定の場合を基準とした提案手法によるトピック依存unigram確率の相対値であり，縦軸の値が1より大きい場合に提案手法によって改善されていることを表している．低頻度語で改善されているケースが多いことから，提案手法では特定のトピックに関連する話題語のunigram確率を引き上げることができていると考えられる．

図4は評価セットslowを構成する各記事（444記事）に対して記事ごとのPPをプロットした結果である．横軸は単一LDA（M=1）で文脈長固定の場合の各記事のPP，縦軸は単一LDAで文脈長固定の場合を基準とした提案手法による各記事のPPの相対値であり，縦軸の値が1より小さいときPPを削減できていることを表す．

提案手法では単一LDA（M=1）のときにはPPが増加する

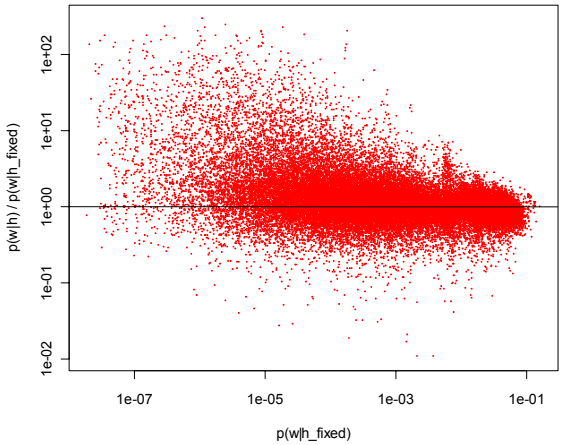

図3　提案手法によるトピック依存 unigram

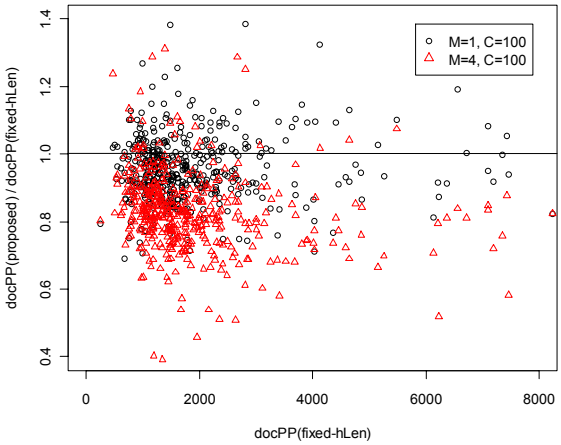

図4　各記事に対するパープレキシティ

記事も少なくないものの，4個のLDA統合時（M=4）ではほとんどの記事でPPを削減できている．提案手法でPPが増加した記事の数はM=1のとき135記事（30.4%）に対して，M=4では33記事（7.4%）であった．また，文脈長固定時のPPが極端に大きい（＞4000）一部の記事を除いて図の左半分のみに着目すると，文脈長固定時のPPが大きい（すなわち予測が難しい）記事ほどモデル統合によるPP削減の度合が大きい傾向がある．

6. おわりに

テキスト入力支援をはじめとするオンラインアプリケーションにおける入力語予測の高精度化を目的として，独立に学習した複数のLDAで文脈長の最適化とトピック適応を行い，この結果を統合する手法を提案した．

提案手法では，複数 LDA の組み合わせによるトピック適応精度向上と複数文脈長からの予測統合によるトピック変化点検出安定化の二つの効果が得られる．文脈長の最適化においては，現在の文の既知部分に対する予測精度に基づいてヒストリ始点候補を評価することによって，入力予測に適した文脈長をより直接的に推定する．そして LDA から得られるトピック依存 unigram を用いて，unigram-rescaling により N-gram をトピックに適応させる．新聞記事コーパスを対象とした評価実験の結果，文脈長固定の場合と比較して話題語の推定精度を向上できること，ほとんどの対象記事に対して PP を削減でき，予測が難しい記事に対して改善の効果がより大きいことが確かめられた．

今後は，文献[8]等の先行手法との比較や計算コストの評価を行うとともに，文脈長最適化と関連が高いトピックセグメンテーションやこれを用いた特徴語抽出，特にテキストストリームを対象としたリアルタイム処理への応用に取り組む．さらにDM，PLSIなど他のトピックモデルにおいて本方式と同様に文脈長最適化を行い，比較評価を行う予定である．

参考文献

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

[2] D. Blei, A. Y. Ng and M. Jordan, “Latent dirichlet allocation”, *Journal of Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[3] 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.

[4] D. Gildea and T. Hofmann, “Topic-based language models using EM”, *Proc. Eurospeech’99*, pp.2167-2170, 1999.

[5] 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.

[6] 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.

[7] 中村明，速水悟，津田裕亮，松本忠博，池田尚志，“複数モデルの統合によるLDAトピックモデルの高精度化とテキスト入力支援への応用”，情報処理学会論文誌，Vol.50, No.4, pp.1375-1389, 2009.

[8] D. Mochihashi and Y. Matsumoto, “Context as Filtering”, *Proc. NIPS2005*, pp.907-914, 2005.

[9] 中村明，速水悟，津田裕亮，松本忠博，池田尚志，“トピック変化点検出に基づくLDAのオンライン適応における複数モデル統合の効果”，言語処理学会第15回年次大会論文集, pp.909-912, 2009.

[10] T. L. Griffiths, and M. Steyvers, “Finding Scientific Topics”, *PNAS*, Vol.101, pp.5228-5235, 2004.

[11] CD-毎日新聞2005/2006データ集

[12] 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.

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

[14] R.Kneser and H.Ney, “Improved Backing-off for M-gram Language Modeling”, *Proc. ICASSP*, vol.1, pp.181-184, 1995.