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

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

N-gramモデルは単語間の局所的な依存関係をモデル化する言語モデルであり，様々な自然言語処理タスクに応用されている．一方，単語間の大域的な依存関係をモデル化する長距離言語モデルとしてはキャッシュモデルやトリガーモデルが知られている[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関係を単語（対）の形で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．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)[2]，PLSIをベイズ学習に基づき改良したLDA(Latent Dirichlet Allocation)[3]，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)[4]などのモデルが提案されている．

これら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[5]等の補間手法によってN-gramモデル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などへの適用が試みられている[6,7].

トピックモデルを用いて入力テキストの話題変化に逐次，適応していく場合（以下，これをオンライン適応と呼ぶ），文脈として用いる単語列の長さが予測精度に影響する．従来の研究では通常，文書先頭以降の全単語を文脈として用いる[4,5]か，文脈長を一定単語数とする方法[6,8]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，またひとつの文書の中でもトピックが時々刻々と変化する上，その速さも一定ではない．したがってトピック変化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[9]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．

これに対し本稿では，トピック変化点検出に基づいて文脈長を最適化することにより，LDAのオンライン適応における精度を向上することができる方式を提案する．提案方式では，適応によって得られる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することにより，トピック変化点を推定する．さらに，独立に学習した複数のLDAから得られるトピック混合比を併用することにより精度が向上することを示す．

以下，2章でLDAについて概説し，3章で提案方式を説明する．そして4章で評価実験の結果を示し5章でまとめを述べる．

2. LDA(Latent Dirichlet Allocation)

2.1. LDAの概要

LDA[3]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$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$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

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

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[3].

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．

2.2. LDAのオンライン適応

未知の文脈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．

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

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

$\Psi(\gamma)$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γ_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．

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

文脈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10～20形態素，即ち1文程度が望ましいと指摘されているが[6,10]，実際には最適な文脈長はテキスト中でのトピック遷移に依存すると考えられる．したがって，トピック変化に応じて文脈長を動的に最適化できることが望ましい．

3. 提案方式

3.1. トピック変化点の検出

提案方式では，現在地点以前における直近のトピック変化点を検出し，このトピック変化点以降を文脈hとしてトピック適応を行う．トピック変化点を検出する手順を以下に示す．

(1)現在地点より前のL_0形態素を処理対象範囲B_0とし（$L_0 > L$; L=文脈長の下限），B_0から複数のトピック変化点候補箇所を抽出する．（図1(a)）
(2)B_0を各トピック変化点候補箇所で2ブロックに分割し，2ブロック間の類似度を計算する．そして類似度が最も小さくなるトピック変化点候補箇所D_1とその類似度S_1を求める．文書ブロック間の類似度は次節で示すカーネル関数により算出する．（図1(b)）
(3)B_0をD_1で分割し，現在地点に近い側のブロックをB_1とする．（図1(c)）

以降は以下(4)(5)の処理を繰り返す($n \geqq 2$).

(4)ブロックB_{n-1}に対し(2)と同様の手順でトピック変化点候補箇所D_nとその類似度S_nを求める．（図1(d)）
(5)$S_n \geqq S_{n-1}$の場合，D_{n-1}をトピック変化点として処理を終了する．$S_n < S_{n-1}$の場合，B_{n-1}をD_nで分割し現在地点に近い側のブロックをB_nとして，処理を継続する．（図1(e)）

なお，原理的には上記(1)におけるトピック変化点候補はB_0中のすべての単語境界であるが，実際には1文中でトピックが変化することは考えにくいことと計算コストを考慮し，本稿では文境界（句点出現位置）をトピック変化点候補とする．

3.2. 文書ブロック間の類似度

文書ブロックbに対して，正規化されたトピック混合比ベクトル$\boldsymbol{t}$を考える．$\boldsymbol{t}$の各要素t_kは，bを文脈hとしてトピック適応を行って得られた変分パラメータγ_kを正規化した値，即ち $t_k = \gamma_k / \Sigma_k \gamma_k$ と定義する．以下，本稿ではこの正規化されたトピック混合比ベクトルを単にトピック混合比ベクトルと呼ぶ．

隣接する2つの文書ブロックb_i, b_j間の類似度を与えるカーネル関数として以下の4種類を用いる．$\boldsymbol{t}_i$, $\boldsymbol{t}_j$はそれぞれb_i, b_jに対応するトピック混合比ベクトルを表す．

(1)Linearカーネル

$$K_L(\boldsymbol{t}_i, \boldsymbol{t}_j) = \boldsymbol{t}_i \cdot \boldsymbol{t}_j$$

(2)Kullback-Leiblerカーネル[11]

$$K_{KL}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \exp\left\{-a\left(KL(\boldsymbol{t}_i \,\|\, \boldsymbol{t}_j) + KL(\boldsymbol{t}_j \,\|\, \boldsymbol{t}_i)\right)\right\}$$

$$KL(\boldsymbol{t}_i \,\|\, \boldsymbol{t}_j) = -\sum_{k=1}^{C} t_{ik} \ln\left(t_{jk} / t_{ik}\right)$$

(3)Fisherカーネル

$$K_F(\boldsymbol{t}_i, \boldsymbol{t}_j) = \Phi(\boldsymbol{t}_i) G^{-1}(\boldsymbol{\alpha}) \Phi(\boldsymbol{t}_j)^t$$

$\Phi(\boldsymbol{t})$: LDAのパラメータ$\boldsymbol{\alpha}$に関するFisherスコア[12]

$G^{-1}(\boldsymbol{\alpha})$: Fisher情報行列の逆行列

(4)Latent Dirichletカーネル[13]

$$K_{LD}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \int \mathrm{Dir}(\boldsymbol{t} \mid \boldsymbol{\lambda}_i)^T \mathrm{Dir}(\boldsymbol{t} \mid \boldsymbol{\lambda}_j)^T d\boldsymbol{t} \quad (0 < T < 1)$$

$\mathrm{Dir}(\boldsymbol{t} \mid \boldsymbol{\lambda}_i)$ ： 文書ブロックb_iのトピック混合比の分布を与える潜在的なトピック分布†

† この分布のパラメータ $\boldsymbol{\lambda}_i$ は変分ベイズ EM アルゴリズムで求められる[9].

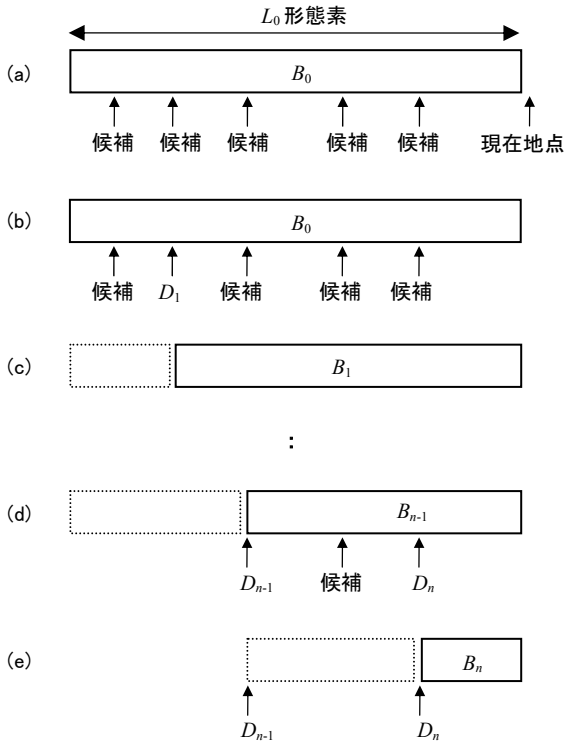

図1．トピック変化点検出手順

隣接する2ブロックの話題が大きく異なっていると類似度が小さな値となり，2ブロックの境界がトピック変化点として検出されやすくなる．

3.3. 複数LDAの統合

LDAでは，複数のモデルで推定した単語生起確率の平均をとることによって予測精度を大きく向上・安定化できることが確かめられている[8]．本稿では，以下の方法によりトピック変化点検出に基づく文脈長の最適化に複数LDAを導入し高精度化を図る．

(1) 独立に学習したM個のLDAモデル(Q_1, Q_2,…,Q_M)を用いて，3.1節のアルゴリズムにより直前のL_0形態素からのべM個のトピック変化点$D^{(m)}$ ($1 \leqq m \leqq M$)を検出する．
(2) 各トピック変化点を始点とするのべM個の形態素列$h^{(m)}$ ($1 \leqq m \leqq M$)を文脈として，各LDAモデル$Q_{m'}$($1 \leqq m' \leqq M$)により語wのトピック依存unigram確率$p_{m'}(w \mid h^{(m)})$を求める．
(3) 前項で得られた計($M \times M$)個の確率の平均を語wの生起確率とする．即ち

$$\bar{p}(w \mid h) = \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} p_{m'}(w \mid h^{(m)})$$

3.1節のアルゴリズムのみでは1個のトピック変化点が確定的に検出されるが，上記の方式では複数個のトピック変化点が検出される．各変化点に基づく推定結果を統合することにより高精度化・安定化が期待できる．

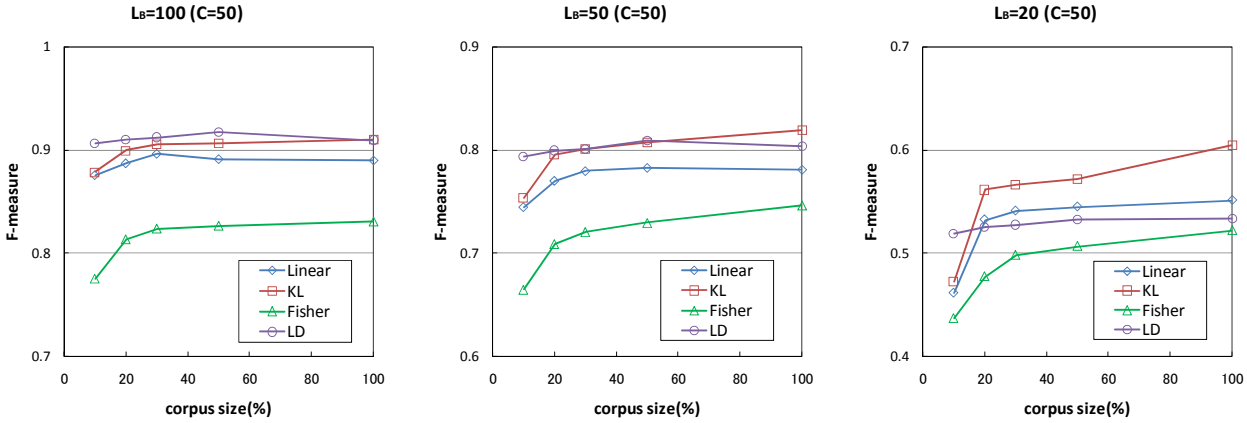

図 2. 記事境界検出によるカーネル関数の比較

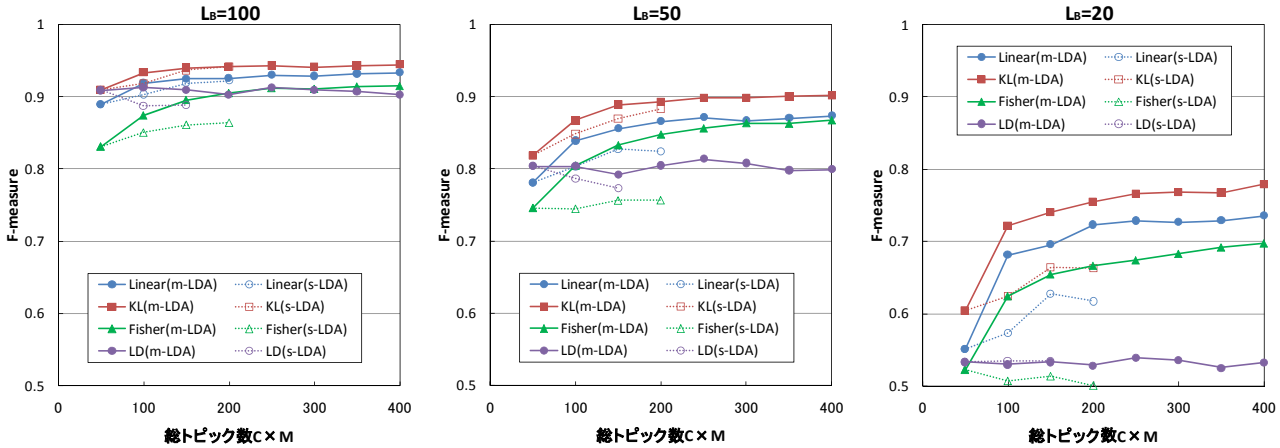

図 3. 記事境界検出における複数 LDA 統合の効果

4. 評価実験

4.1. 学習テキストおよび評価テキスト

学習テキストおよび評価テキストを以下に示す.

[学習テキスト]
CD－毎日新聞2005データ集[14]全記事(95881件). のべ約2864万形態素, 異なり語数185196

[評価テキスト]
CD－毎日新聞2006データ集[14]の内, 200文字以上の記事から無作為抽出した1000件の記事(のべ約40万形態素)を連結し1文書とした.

学習テキスト・評価テキストとも, 文節構造解析システムibukiC[15]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1746語であった.

4.2. 学習条件

LDAの学習では学習テキスト中の出現回数が4回以下の語を除いた77794語を語彙とした. 収束判定は, 学習テキストに対する1ステップ前からのパープレキシティの減少が0.1%未満となった時点で収束とした. 3.3節で述べた複数LDAを用いる方式では, α, βに異なる初期値を与えて同一の学習テキストに対しM回の学習を行い, M個のLDAを構築した.

4.3. 記事境界検出によるカーネル関数の評価

トピック変化点検出に基づくLDAのオンライン適応に先立ち, 3.2節で示した4種類のカーネル関数を比較評価するため下記の方法で記事境界検出実験を行った.

(1)1000記事を連結した評価テキストに対して, 各記事の先頭から文単位で形態素数をカウントし, 所定の形態素数下限L_B(=100,50,20)に達したor超えた文の文末で分割
(2)各記事の末尾で生じるL_B形態素より短い部分は直前のブロックに含める
(3)各分割位置において前後2ブロック間の類似度を算出し, 類似度がしきい値未満の場合に記事境界として検出

図2に結果を示す. 横軸はLDAの学習に用いたコーパスサイズであり100%が全学習テキストを表す. 縦軸は類似度の検出しきい値を変えてF尺度が最大となった時の値をプロットした. LDAの潜在トピック数Cは50トピック, LD(Latent Dirichlet)カーネルのパラメータTは予備実験により0.1とした. 図より, 極端にコーパスサイズが小さい場合を除いてKL(Kullback-Leibler)カーネルが最も安定している. LDカーネルはコーパスサイズの影響を受けにくいが, ブロック長が短くなると大幅に精度が低下する. 以上より, 1文(約20形態素)程度のブロック単位でトピック変化を最も精度良く検出できるのはKLカーネルであるといえる.

KL情報量に基づいて設計されたKLカーネルは本来，確率分布間の類似度を測るカーネルであるが，正規化された周波数スペクトルやヒストグラムに対しても適用されており，対象が多峰性のピークを持つ場合にロバストであることが指摘されている[11]．マルチトピックモデルであるLDAではトピック混合比は通常，数個の潜在トピックのみが大きな値を持つためKLカーネルが適しているのではないかと考えられる．

一方，LDカーネルは確率分布間の類似度を与えるBhattacharyyaカーネルの一種であり，原理的に優れた性能を持つと考えられるが，本実験では良い結果が得られなかった．L_B=100ではKLカーネルの性能を上回ることから，トピック変化点検出よりも文書分類・検索などある程度長いテキスト間の類似度を測る用途に向いている可能性がある．

次に，記事境界検出において複数LDA統合の効果を調べた結果を図3に示す．m-LDAは潜在トピック数Cを50トピックとしてM(≦8)個のLDAを統合した場合，s-LDAは単一LDA（即ちM=1）で潜在トピック数Cを変化させた場合を表す．m-LDAでは各LDAにより得られたM個の類似度の平均値により記事境界を検出した．図より，m-LDAではLDカーネルを除くすべてのカーネルで検出精度が向上している．特にブロック長が短い時ほど有効である．以上より複数LDAの統合がトピック変化点検出においても有効であることが期待できる．

4.4. トピック変化点検出に基づくオンライン適応

3.1節および3.3節で述べたアルゴリズムによりトピック変化点検出を行いオンライン適応を行った結果を図4に示す．**m-LDA**の場合の潜在トピック数Cは前節と同様50トピックとし，カーネル関数は前節の実験で最も高精度であったKLカーネルを用いた．予備実験により，処理対象範囲を決めるL_0は100形態素，文脈長の下限Lは5形態素とした．

s-LDA, **m-LDA**ともに，トピック変化点を検出して文脈長を制御することにより文脈長を20形態素に固定した場合よりもパープレキシティが削減されている．**s-LDA**では潜在トピック数を増やしても性能が向上しないのに対して，**m-LDA**ではモデル数の増加とともに性能が向上している．**m-LDA**(M=4)でトピック変化点検出を行った場合，**s-LDA**(C=50)で文脈長固定の場合に対しては14.7%(1332.6 → 1137.1)，**s-LDA**(C=50)でトピック変化点検出を行った場合に対しては11.5%(1284.9→1137.1)パープレキシティを削減できている．

5. まとめ

トピックモデルLDAのオンライン適応における予測精度を向上することを目的として，トピック変化点検出に基づいて文脈長を動的に制御する方式を提案，評価を行った．提案方式では，適応によって得られる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することにより，トピック変化点を推定する．

新聞記事コーパスを用いた実験の結果，提案方式では文脈長を固定とする場合よりもパープレキシティを削減できること，独立に学習した複数のLDAによる推定結果を組み合わせることによりさらに予測精度を向上できることが確かめられた．隣接する文書ブロック間の類似度を与えるカーネル関数を比較評価した実験では，KL情報量に基づくKLカーネルが最も良好な結果を示した．

今後は本手法をunigram rescaling等の補間手法を用いたN-gramのトピック適応に適用するとともに，Particle Filterによる文脈長最適化手法[9]など他手法との比較評価を行う．

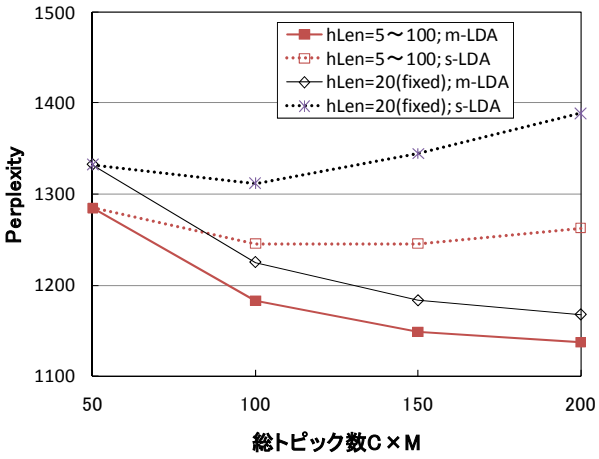

図 4. トピック変化点検出に基づくオンライン適応

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

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

[4] 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.

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

[6] 高橋力矢，峯松信明，広瀬啓吉，“文脈適応による複数N-gramの動的補間を用いた言語モデル”, 情報処理学会研究報告SLP-46-7, pp.37-42, 2003.

[7] 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.

[8] 中村明，津田裕亮，松本忠博，池田尚志，速水悟，“複数モデルの統合によるLDAトピックモデルの高精度化”, 言語処理学会第14回年次大会論文集, pp.305-308, 2008.

[9] 持橋大地，松本裕治，“Particle Filterによる文脈の動的ベイズ推定”, 情報処理学会研究報告2005-NL-165, pp.59-66, 2005.

[10] 津田裕亮，中村明，松本忠博，池田尚志，“LDAトピックモデルにおける文脈推定精度と文脈長に関する考察”, 言語処理学会第14回年次大会論文集, pp.623-626, 2008.

[11] 石垣司，樋口知之，渡辺嘉二郎，“Kullback-Leiblerカーネルによる正規化周波数スペクトル判別とその圧力調整器劣化診断への応用”, 電子情報通信学会論文誌D Vol.J90-D, No.10, pp.2787-2797, 2007

[12] G. Chandalia and M. Beal, “Using Fisher Kernels from Topic Models for Dimensionality Reduction”, *NIPS 2006 Workshop on Novel Applications of Dimensionality Reduction*, 2006.

[13] D. Mochihashi, “Latent Dirichlet kernel & Bayesian kernels”, *ÜberSVM2004*, 2004.

[14] CD-毎日新聞2005/2006データ集

[15] 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.