LDA トピックモデルに基づく話題変化点検出

津田裕亮[1)]，中村明[2)]，速水悟[1)]，松本忠博[1)]，池田尚志[1)]
岐阜大学工学部[1)]，三洋電機（株）エコロジー技術研究所[2)]

1　はじめに

確率・統計的自然言語処理や音声認識の分野で広く用いられている N-gram モデル [1] は，単語の生起を直前の $(N-1)$ 単語を用いてモデル化したものである．単純ではあるが非常に強力なモデルとして知られており，テキスト入力や音声認識などに応用されている．また，単語間の大域的な依存関係を単語対の関係でモデル化し，N-gram モデルと組み合わせたモデルに，トリガーモデルやキャッシュモデルなどがある．これに対し，単語間の大域的な依存関係を話題（トピック，文脈）としてモデル化したものとしてトピックモデルがあり，PLSI（Probabilistic Latent Semantic Indexing）[2] や LDA（Latent Dirichlet Allocation）[3]，DM（Dirichlet Mixtures）[4] など，さまざまなモデルが提案されている．トピックモデルは，現在の話題に応じて単語の生起確率を動的に推定でき，言語モデルの高精度化が期待できる．

トピックモデルを用いた研究において，トピック推定に用いる形態素列（ヒストリ）の長さは単純に固定長とする場合が多い．しかし，ヒストリを固定長にすると，ヒストリの途中でトピックが大きく変化する場合に，単語の推定精度低下を招いてしまう恐れがある．

この問題に対処するため，我々は昨年に，LDA モデルにおけるヒストリ長の特徴を実際の事例を基に調べた [5]．その結果，長さが異なる複数のヒストリを統合することで固定長ヒストリより精度が若干向上することや，推定する単語の品詞とヒストリの長さに関連があることがわかった．しかし，決定的な解決法を見つけることはできず，さらに有用な方法を考えていく必要があった．

そこで本稿では，新たな試みとして，LDA モデルの適応で得られるトピック混合比に着目し，トピック変化点を検出する方法を検討する．隣接ブロックのトピック混合比の距離により変化点を検出しヒストリを制御，固定長ヒストリとの精度比較を行った．

以下，2 章で LDA の概要を述べ，3 章で固定長ヒストリの特徴と問題について述べる．そして 4 章で本手法の概要を説明し，5 章で評価実験の結果と考察を示す．最後に 6 章でまとめと今後の展望を述べる．

2　LDA の概要

LDA（Latent Dirichlet Allocation）[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 $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_C)$ が多項分布の共役事前分布であるディリクレ分布 $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})$ 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）．

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

$\boldsymbol{\alpha}$,$\boldsymbol{\beta}$ 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 uni-gram 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

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

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]．未知のヒストリ 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される．即ち，h に対する変分パラメータ 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．

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

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

$\Psi(\gamma)$ 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 をヒストリ h の元での各潜在トピックの混合比とする．したがって，ヒストリ h の元での語 $w_{j'}$ の生起確率は次式により与えられる．

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

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　固定長ヒストリの精度について

図 1 に，ヒストリの長さ (L_h)，潜在トピック数の違いによる TPP（テストセット・パープレキシティ）の推移を示す（詳しくは [5] 参照）．図 1 より，ヒストリの長さは固定長の場合，計算コストを考慮して 20 程度が妥当と言える．

しかし，実際の文書では，トピックは緩やかに変化する場合もあるし，急激に変化することもある．トピックの変化を動的にとらえ，現在のトピックに適したヒストリを与えることで，単語の推定をより正確に行うことができると考えられる．

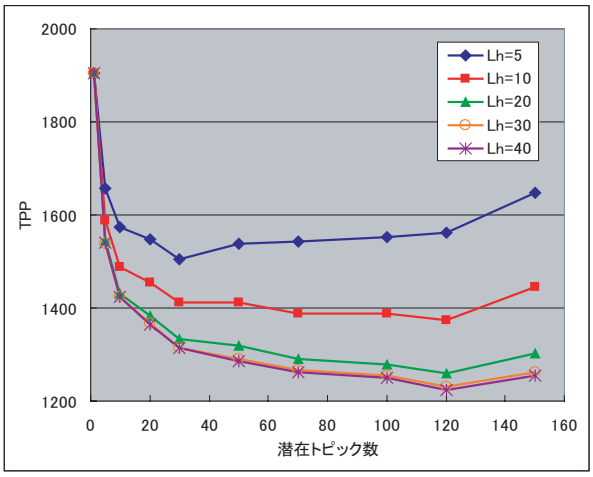


図 1:ヒストリの長さと潜在トピック数の違いによる TPP の推移

4　提案手法

4.1　概要

LDA は，一つの文書には複数の潜在トピックが同時に混在していると考えたモデルであり，その状態はトピック混合比によって確認することができる．トピック混合比は潜在トピック数を次元数とするベクトルで表される $(\gamma_1, \gamma_2, ..., \gamma_C)(C$: 潜在トピック数). 図 2 は，文単位でトピック混合比ベクトルの遷移を表したイメージ図である．

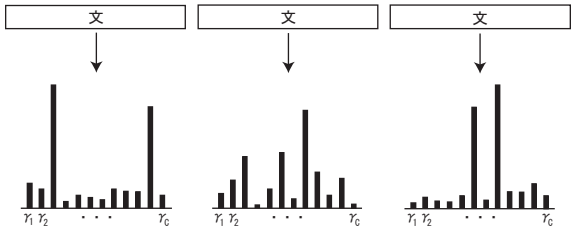
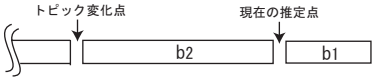


図 2:トピック混合比ベクトルの遷移の様子

本手法では，トピックの変化点候補を文末と仮定し，各文のトピック混合比ベクトルの変化量によりトピック変化点かどうかを判定する．

4.2　手順

手順を以下のような図を用いて説明する．



現在の文のトピックが，その前のトピックと同じであるかどうかを判定してヒストリ h を決定する．現在の文の単語列をブロック b_1，過去にトピック変化点と判定された点から b_1 直前までの単語列をブロック b_2 とする．よって，b_1 と b_2 の間が現在の判定を行う点ということになる．

判定を行う前に，まず，b_1,b_2 のトピック混合比ベクトルを合計 1 となるように正規化を行う．正規化されたトピック混合比ベクトル$\boldsymbol{t_1}$,$\boldsymbol{t_2}$間の距離を，以下の距離尺度により算出して判定を行う．

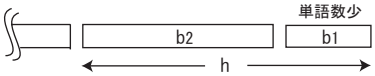
- ユークリッド 2 乗距離
$$D^2(\boldsymbol{t_1}, \boldsymbol{t_2}) = ||\boldsymbol{t_1} - \boldsymbol{t_2}||^2$$

- 対称化 KL ダイバージェンス
$$sKL(\boldsymbol{t_1}||\boldsymbol{t_2}) = KL(\boldsymbol{t_1}||\boldsymbol{t_2}) + KL(\boldsymbol{t_2}||\boldsymbol{t_1})$$

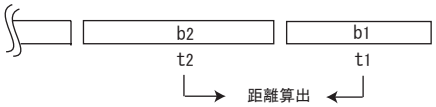
- Jensen-Shannon ダイバージェンス
$$JS(\boldsymbol{t_1}||\boldsymbol{t_2}) = \{KL(\boldsymbol{t_1}||\boldsymbol{t_{12}}) + KL(\boldsymbol{t_2}||\boldsymbol{t_{12}})\}/2$$
$$\boldsymbol{t_{12}} = (\boldsymbol{t_1} + \boldsymbol{t_2})/2$$
$$KL(\boldsymbol{t_i}||\boldsymbol{t_j}) = -\sum_k t_{ik} ln \frac{t_{jk}}{t_{ik}} \quad (t_{ik}:\boldsymbol{t_i}\text{の第 } k \text{ 要素を表す})$$

以下に，具体的手順を示す．

- b_1 の単語が一定数溜まるまでは，$h = b_1 + b_2$ とする．

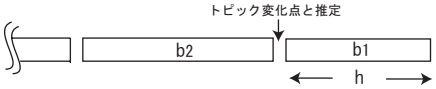


- b_1 の単語が一定数溜まったら，$\boldsymbol{t_1}, \boldsymbol{t_2}$ の距離を算出し，閾値以上の場合に変化点と判定する．



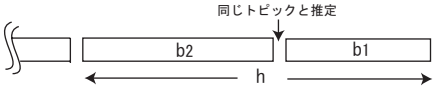
＜変化点と判定された場合＞

$h = b_1$ とする．



＜同じトピックと判定された場合＞

$h = b_1 + b_2$ とする．



次の文になるまでこの手順を繰り返し，次の文になった時点で，今まで判定してきた点を最終的にトピック変化点であるかどうかを決定する．また，ヒストリが長くなりすぎると精度低下の原因になることがわかっているため，ヒストリ長に上限を設ける．

5 評価実験

距離尺度別に変化点判定の閾値を変えながらモデルを評価する．評価は，固定長の最適値だと思われるヒストリ長 20 の TPP(uni-gram) と比較して行う．モデルの潜在トピック数を 50，b_1 の必要最低限単語数を 5，ヒストリの最大単語数を 100 に設定した．

5.1 実験データ

学習文書，評価文書は毎日新聞 [6] の記事を使用し，形態素解析は我々の研究室で開発をしている ibukiC [7] で行なった．

【学習文書】
毎日新聞 2005 年の記事番号奇数の記事
48035 記事，異なり 141666，延べ約 1439 万形態素
頻度 2 以下の語を除いた 75314 語でモデル構築*

【評価文書】
毎日新聞 2005 年の記事番号偶数の記事
ただし，200 文字以上の記事としている
次節で述べる方法により複数作成

5.2 評価文書の作成

提案手法は，トピックの動的な変化をとらえるモデルである．したがって，いつも同じトピックが続く文書ではなく，トピックが様々な速度で変化するような文書が好ましい．そこで [8] にならい，トピックの変化速度が異なる 3 種類の評価文書を作成する．手順は以下の通りである．

記事カテゴリ†別に複数の文書を用意し，それをサブテキストとする．

(1) 各サブテキストに対して文の読み取り位置を設ける．
(2) 無作為に一つのサブテキストを選び，連続する X 文を採取する．
(3) (2) の後に，読み取り位置をさらに Y 文進める．
(4) 必要な文数が得られるまで (2)(3) を繰り返す．

X, Y は表 1 に従う乱数．今回の実験では，1000 文採取したものを変化速度別に 3 セットずつ用意した．

表 1：評価文書の種類

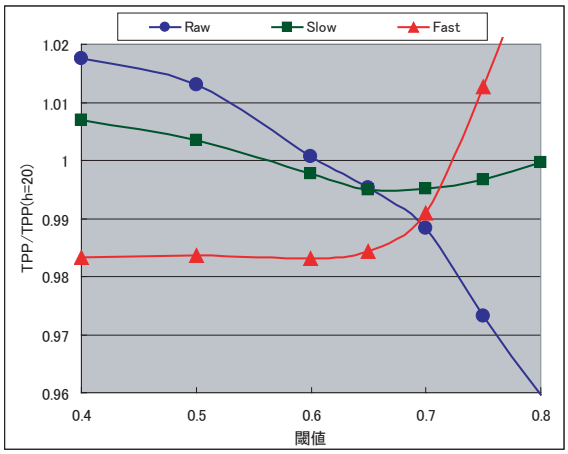
Name	Property
Raw	X=100,Y=0
Slow	$1 \leqq X \leqq 10$,$1 \leqq Y \leqq 10$
Fast	$1 \leqq X \leqq 3$,$1 \leqq Y \leqq 10$

*学習時の収束判定は，全学習文書に対する 1 ステップ前からのパープレキシティの減少が 0.5 未満となった時点で収束とした．

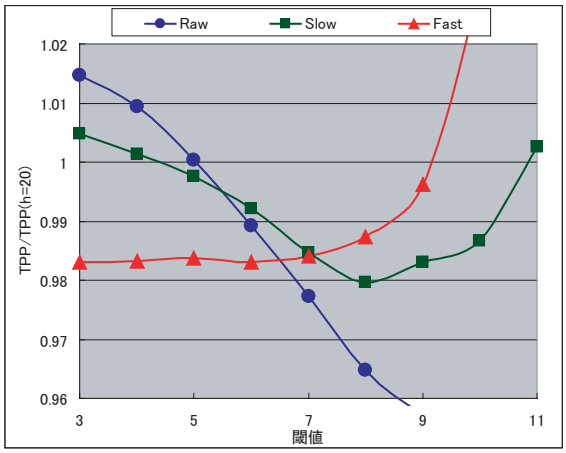
†1 面　2 面　3 面　解説　社説　国際　経済　特集　総合　家庭　文化　読書　科学　芸能　スポーツ　社会

5.3 実験結果

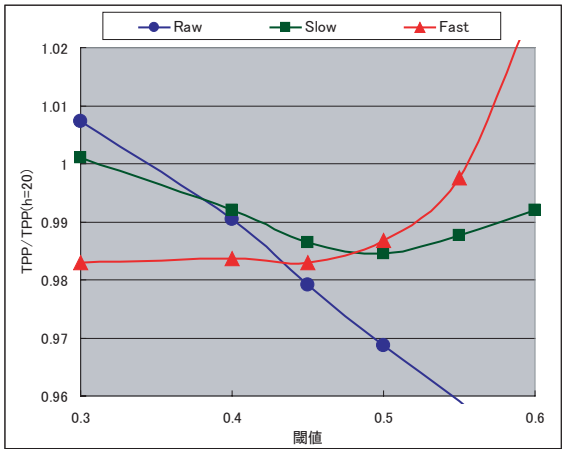
図 3 に，各評価文書の結果を距離尺度別に示す．各グラフの横軸がトピック変化点の判定をする際の距離尺度の閾値を表し，縦軸が本手法の TPP とヒストリ長を 20 で固定した場合の TPP($\mathrm{TPP}_{h=20}$) との比率 ($\mathrm{TPP}/\mathrm{TPP}_{h=20}$) を表す（3 セットの平均）．



(a) ユークリッド 2 乗距離



(b) 対称化 KL ダイバージェンス



(c) Jensen-Shannon ダイバージェンス

図 3:距離尺度別実験結果

本手法では，トピックの変化点候補を文末としているため，Fast のように一つのトピックの長さが短い文書は，距離尺度の閾値が小さいほど精度が良い結果となっていた（単純に文末を変化点とした場合でも精度が良くなっている）．また，Raw は一つのトピックの長さが長いため，同じトピックと判定されやすいほど精度が良くなる傾向にあった．実際はトピックの変化速度は未知であるため，適正な閾値は Slow で判断するのが妥当であると考えられる．

距離尺度は，ユークリッド 2 乗距離においては効果はわずかであるが，他の 2 種の方法，特に対称化 KL ダイバージェンスが安定して精度が良くなっていた．

5.4 誤り事例

図 4 に，b_1 が本来は前の文と同じトピックの文であるのに，トピック変化点と判定されてしまった一例を挙げる．図 4 の上部の実線の囲いが b_2，下部の実線の囲いが b_1 であり，さらに b_2 では同一記事カテゴリの文を破線で囲っている．

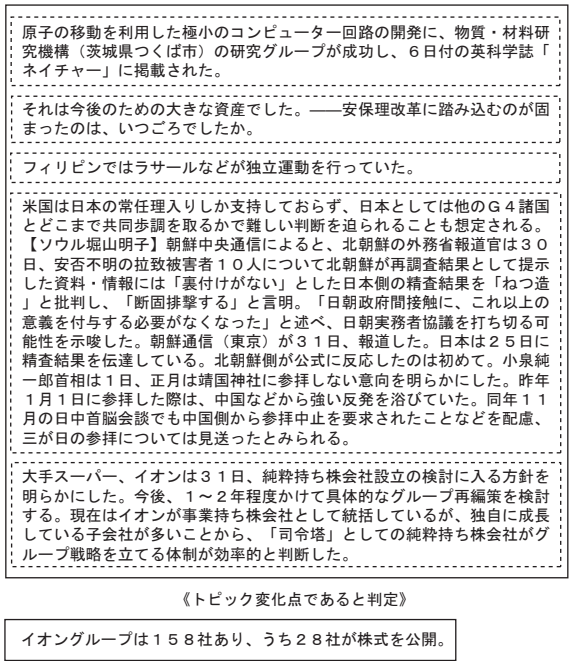


図 4:トピック変化点と誤判定されてしまった例

本来，b_2 の各文は同じトピックの文でなければならないのだが，実際は複数の記事カテゴリが混在しており，過去のトピック変化点の判定に失敗していることがわかる．失敗の原因の一つとして、比較するブロックが極端に短い場合が考えられる．比較ブロックが短すぎると，信頼性のあるトピック混合比が得られない．しかも本手法では，b_1 は常に短い（最大でも 1 文）ため，判定に失敗する可能性が高くなってしまう．このようなことが原因で，図 4 の例のように，b_2 に複数のトピックが混在してしまい，本来は繋がるはずの文が繋がらないという結果を招いてしまっている．

この問題を解決する手段の一つとして，b_2 の長さに上限を設けるという方法が考えられる．これは，b_2 の長さが長いほど，別のトピックの内容が含まれてくる可能性が高くなるのではないか，という考えに基づいている．また，過去に判定された点の判定し直しという方法も考えられる．本手法では，一度トピック変化点であるかそうでないかを判定したら，その後の判定し直しは行っていない．このため，過去の失敗が原因で現在の判定に影響が出てしまっている可能性がある．再度判定を行うことで，過去の失敗の影響を排除することができるのではないかと考えられる．

6 おわりに

本稿では，LDA モデルの適応で得られるトピック混合比によりトピック変化点を判定する方法を提案した．3 種の距離尺度によりトピック変化点を判定，固定長と比較した結果，評価文書で違いはあるものの，TPP が約 1.3～3.5 %低下した（対称化 KL ダイバージェンスの閾値 8 の場合）．

しかし，今回の方法は非常に単純であるため，改善すべき点は多い．今後は，トピック変化点の再判定，複数モデルで判定を行うことによる判定の安定化などを試みていく予定である．

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. of 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] 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.

[6] CD-毎日新聞 2005 データ集

[7] 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.

[8] 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, 2003-NL-155, pp.107-112, 2003.