Particle Filter による文脈の動的ベイズ推定

持橋 大地 [1,2]　松本 裕治 [1]
[1] 奈良先端科学技術大学院大学 情報科学研究科
[2] ATR 音声言語コミュニケーション研究所 音声言語処理研究室
{daiti-m,matsu}@is.naist.jp

概要

文脈をとらえる長距離言語モデルの研究において, これまで, 必要な文脈長の問題はあまり議論されることがなく, 単純に文書の先頭から用いるなどの方法が行われてきた. 本論文はこれに対し, 文脈の変化に対する明示的な確率的生成モデルを与え, 話題の変化とその速度をとらえ, 必要な文脈長を自動的に選択することのできるベイズ言語モデルを提案する. 提案法は TextTiling の確率化ともとらえることができ, 非線型フィルタである Particle Filter によって解かれる. BNC コーパスでの実験により, 単純な履歴を用いる従来のベイズ言語モデルに対して, 高い性能を示した.
キーワード: Particle Filter, Mean shift model, 変化点検出, 時系列モデル, 長距離言語モデル

A Particle Filter approach to dynamic Bayesian context estimation

[1,2]Daichi Mochihashi　[2]Yuji Matsumoto
[1]Graduate School of Information Science, NAIST
[2]ATR Spoken Language Translation Research Laboratories
{daiti-m,matsu}@is.naist.jp

Abstract

This paper proposes a novel Bayesian long-distance language model that can capture subtopic shifts within a document. To model these subtopic flows, we introduce a latent mean shift model of natural language, and estimate its state space by a Particle Filter. Experiments on BNC corpus showed consistent improvements over the naïve context model that has been used so far.
Key words: Particle Filter, Mean shift model, Change point analysis, Time series analysis, Long-distance language model

1　はじめに

「文脈」をとらえることは言語活動の基本的な要素であり, われわれはその場の文脈を判断し, 適切にモデルを切り替えてゆくことで, 適応的に言語理解や発話を行っている. たとえ句読点の一切ない小説 [18] であっても, そこには明確に話の文脈が存在し, むしろ文脈の流れをとらえていくこと自体が, テキストを読むということの大きな要素の一つだと言ってよい.

工学的にみても, 単一の文脈に適応するだけでなく, 複数の文脈や状況の変化をとらえ, 動的に適応してゆくことは, 実時間の連続音声認識やロボティクスなどにおいて特に重要であると考えられる.

自然言語処理においては, これは n-グラムより高次の関係をとらえる, 長距離言語モデルの適応化の問題と考えることができる. 短距離の文法的な確率は文脈の影響を受けにくいが, 単語のユニグラム出現確率のような意味的な確率は文脈の影響をきわめて大きく受けるため, 文脈への適応化は大きな課題である.

長距離文脈をとらえる問題は, キャッシュやトリガのような古典的なモデル [14] から始まり, LSI を用いて直接共起しない関係を考慮することのできる言語モデル [2] を経て, 近年では隠れ変数を用いた混合モデルの推定問題として定式化され, 確率的言語モデルとの相性がよく, 従来法より高い予測性能をもつことが報告されている. [10, 21, 22, 23]

しかしながら, それらのモデルにおいて, 必要な適切な文脈長の問題はほとんど議論されてこなかった. これらのモデルは基本的にテキストモデルの応用であり, 履歴の時間的な順序を考慮せず, Bag of Words として捉えるものである. このため, 履歴としては文書の最初から全てを用いるか [10, 21, 22], 1000 単語前までなどの単純な閾値を用いる [15] ことが行われてきた.

しなし, これはあくまでも近似であり, 実際のテキストがそうなっているわけではない. TextTiling [12] はこのような不均質性に従ってテキストをサブト

ピックに分割するアルゴリズムであるし, Beefermanら [1] は同様に言語の時間的な非定常性に着目し, セルフトリガ (同じ語の再出現) の分布から, テキスト中での語の意味的な関係がテキスト中での間隔に従って指数的に減少することを見出している.

別の言葉で言えば, 今までの確率的なテキストモデル, およびそれに基づく言語モデルは, テキストがどれほど長くても, 1 つの定常情報源から生まれたと仮定し, そのパラメータを順次精密に求めるアプローチであることを意味する.[1]

本論文ではこれに対し, 文脈の変化に対する明示的な確率モデルを与え, そのパラメータをオンラインで時系列に従って推定することにより, 話題の変化とその速度をとらえ, 適切な文脈長を自動的に選択することのできるベイズ言語モデルを提案する.

このモデルは非線型な HMM であり, 従来の Baum-Welch 法やカルマンフィルタ等では解くことができないが, 近年計算量的に利用可能となってきた, モンテカルロ法を用いたベイジアンフィルタである Particle Filter を用いることで解くことができる.

2 章で, Mean Shift Model と呼ばれるこのためのモデルについて述べ, 3 章で Particle Filter について説明する. 4 章で Mean Shift Model を自然言語に拡張し, 確率的なテキストモデルである DM および LDA を用いた MSM-DM, MSM-LDA を導入する. 5 章で BNC を用いた実験結果と考察を示し, 6 章でまとめとこれからの展望について述べる.

2 Mean Shift Model

はじめに述べた確率的文脈モデルはいずれも, 文脈には隠れたユニグラム分布, あるいは確率的トピック分布という多項分布が存在すると仮定し, 入力履歴に従ってその推定値を更新することで, 次の語の予測を行うモデルである.

したがって, 文脈追跡のためには, 隠れた多項分布自体の変化をとらえるモデルが必要になる. このためのモデルの一つ[2] が Mean shift model (MSM) である.

これは HMM の一種であるが, 通常の離散 HMM とは違うことに注意したい. 通常の HMM では, 真の状態は M 個の離散状態のどれか一つであり, その確率的な推定値として多項分布を得るが, ここでは, 真の状態自体が多項分布であり, その確率的な推定値として多項分布の分布 (ディリクレ分布または混合ディリクレ分布) を得ることになる.

離散変数上への分布自体を状態とする HMM という意味で, これは Ghahramani らの Factorial HMM (Ghahramani 1995) に似ているが, FHMM のようにダイナミクスのパターンを固定するのではなく, テキストによってパターンの一つ一つ異なるランダムウォークを追跡することを目的としている.

Blei らは, PLSI[13] の事後多項分布を, 確率の最も高い一点で近似することで離散 HMM を構成し, 異なるテキストの境界を検出する Aspect Hidden Markov Model を提案している [3]. しかし, 違ったテキストの境界ではなく, テキスト内部のサブトピックの変化をとらえるためには, 2 番目以降[3] の山の変化が重要であり, 多項分布の変化を直接モデル化する必要がある. この意味で, 本研究は [3] の厳密化であるともいえる.

以下で, 多項分布の Mean shift model について説明する.

2.1 Multinomial Mean shift model

Mean shift model (MSM) とは, 隠れ状態の間欠的な変化を記述する生成モデルであり, 正規分布について導入されたものを [7][19], 近年 Chen and Lai [6] により, Particle Filter を用いることで変化率をも動的に推定する拡張がなされたが, 紙面の都合上省略し, [6] での, DNA 分析における多項分布に対する拡張についてのみ以下で説明する.

多項分布の MSM では, 観測されたアルファベット系列 $\mathbf{y} = (y_1 y_2 \dots y_T)$ ($y_t \in \mathcal{A}$ は離散アルファベット集合) を出力した真の多項分布 θ が複数存在し, 時間的に変化していると考え, 次のような生成モデルを仮定する.

$$\begin{cases} \theta_t \sim \mathrm{Dir}(\alpha) & \text{with probability } \rho \\ \quad = \theta_{t-1} & \text{with probability } (1-\rho) \\ y_t \sim \mathrm{Mult}(\theta_t) & \end{cases} \quad (1)$$

ここで $\mathrm{Dir}(\alpha)$, $\mathrm{Mult}(\theta)$ はそれぞれ, α, θ をパラメータにもつディリクレ分布および多項分布である.

このモデルでは, 最初に多項パラメータ θ を $\mathrm{Dir}(\alpha)$ からサンプルし, しばらくの間 θ から y を出力する. 確率 ρ で文脈の変化が起こると, また新しい θ が $\mathrm{Dir}(\alpha)$ からサンプルされ, 以後の y はそこから出力する. このプロセスを繰り返す. 以上において, θ はもちろん, 変化点の場所もわれわれには未知であり, 観測されるのは出力系列 $\mathbf{y}$ のみである.

例として, 図 1 の $T = 100$ の系列を考える. ここでは, アルファベットは $\mathcal{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ である. この系列において, 次の出力 y は何であろうか.

[1] これは従来のテキストモデルが, 新聞記事のような比較的均質で短いテキストを学習データとして用いていたことにも依っていると思われる. 長い, 構造的なテキストの標準的なコーパスは驚くほど少ない. [12]

[2] 本研究に先立ち, 文脈に対して一様なブラウン運動を仮定し, Power steady model (Smith 1979) に基づいて事前分布を確率 $\gamma \sim \mathrm{Be}(a, b)$ で忘却し, γ のもつベータ分布のハイパーパラメータをオンラインのカーネル密度推定 [9] で求めるアプローチを行ったが, あまり良い結果を得られなかった.

[3] トピックは一般に数百存在するため, 点推定による近似は, 非常に粗い近似となる可能性が高い.

```
aaaaaabaacbaabaaaaabbbbbababababaaba\
babbabbbbabcaccccbcacaccccccccccccc\
ccaccccacccccccccccccacaaaacbbbbb
```

図 1: 観測された時系列データ.

明らかに, この推定値は直前の変化点がどこであるかに依存する. いま, 時間 t において変化が起こったかどうかを表す二値変数を I_t としよう. $I_t = 1$ は時間 t において変化が起こった ($\theta_t \neq \theta_{t-1}$) ことを, $I_t = 0$ は変化が起こらなかった ($\theta_t = \theta_{t-1}$) ことを意味する.

$I_t = 1$ の場合:

この場合, 図 2(a) のように, 時刻 t において変化が起こり, $\theta_t \sim \mathrm{Dir}(\alpha)$ が新しくサンプルされ, そこから y が出力されたのであるから, その確率は

$$p(y|Y_{t-1}, I_t = 1) = \int p(y|\theta_t)p(\theta_t|\alpha)d\theta_t \quad (2)$$

$$= \alpha_y / \left(\sum_{i=1}^{|\mathcal{A}|} \alpha_i \right) \quad (3)$$

となる.

$I_t = 0$ の場合:

この場合, 最近の変化点を $t = c$ とすれば ($I_c = 1, I_{c+1} = \cdots = I_{t-1} = 0$), 図 2(b) のように, 時刻 c において $\theta_c \sim \mathrm{Dir}(\alpha)$ がサンプルされ, $y_c \cdots y_{t-1}$ を出力した後に y が出力されたのだから, その推定値は

$$p(y|Y_{t-1}, I_t{=}0) = \int p(y|\theta_t)p(\theta_t|y_c \cdots y_{t-1})d\theta_t \quad (4)$$

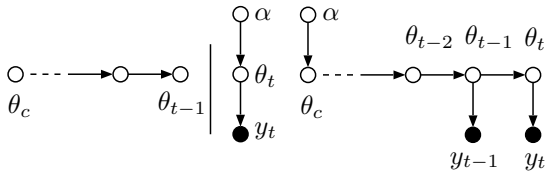
$$= \int \theta_y \cdot \mathrm{Dir}(\alpha + \sum_{t=c}^{t-1} \delta(y_t))d\theta_t \quad (5)$$

$$= \frac{\alpha_y + \sum_{t=c}^{t-1} \delta(y)}{\alpha + \sum_{t=c}^{t-1} \delta(y_t)} \quad (6)$$

と求まる. ここで, $\delta(y)$ は y に確率密度が集中する Dirac の δ 関数.

このように, 直前の文脈の変化点がわかっていた場合, 予測分布は閉じた形で求まるため, 変化点を求めることがこの問題の本質であることがわかる. これは統計学において, 変化点検出問題 [17] として知られている問題の一種である. 下に述べるように, 直前の変化点の位置は, その 1 つ前の変化点の位置に依存する. 同様にして再帰的な依存関係があるため, この問題を解くには, 少なくとも非線型な動的計画法が必要となる.

上式において, 変化点 $t = c$ は計算上確定されなければならないため, 安定した推定を行う方法として, オンラインのモンテカルロ法である Particle Filter が有用である. 次節で Particle Filter について簡単に説明し, 上の問題のオンライン推定法を述べる.



(a) $I_t = 1$ の場合　　(b) $I_t = 0$ の場合

図 2: Mean shift のグラフィカルモデル.

3 Particle Filter

3.1 Particle Filter と重点サンプリング

Particle Filter [9] とは, モンテカルロ法をオンラインで行うアルゴリズムであり, 近年の計算資源の増大に伴い, 主に実ベクトル空間を対象として, 信号処理やロボティクスなどの分野で使用されてきた. 重点サンプリング法 [11] を時系列的に行うものと考えられるため, SISR(Sequential Importance Sampling/Resampling) とも呼ばれている.

従来のカルマンフィルタやその拡張などと異なり, 線形モデルや正規分布だけでなく, 任意の非線型な分布を追跡することができる. そのため, 本論文のように, 自然言語のような離散データにも原理的に適用可能である.

重点サンプリング (IS) とは, ベイズ推定の期待値計算において, 積分をサンプリングにより近似する方法であり, 確率変数 $\mathbf{x}$ の関数 $f(\mathbf{x})$ の期待値を以下のように近似する.

$$I = \int p(\mathbf{x})f(\mathbf{x})d\mathbf{x} \quad (7)$$

$$= \int q(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}f(\mathbf{x})d\mathbf{x} \quad (8)$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}f(\mathbf{x}^{(i)}) \quad \left(\mathbf{x}^{(i)} \sim q(\mathbf{x})\right) \quad (9)$$

$$= \sum_{i=1}^{N} w(\mathbf{x}^{(i)})f(\mathbf{x}^{(i)}) \quad \left(w(\mathbf{x}^{(i)}) = \frac{1}{N}\frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}\right) \quad (10)$$

ここで, $q(x)$ は $p(x)$ よりサンプリングが容易な分布であり, 提案分布と呼ばれる. 式 (10) から, これは $\mathbf{x}^{(i)} \sim q(\mathbf{x})$ に対し, $f(\mathbf{x}^{(i)})$ を $w(\mathbf{x}^{(i)})$ で重みづけて和をとることで, $f(\mathbf{x})$ の期待値 $E[f(\mathbf{x})]$ が求まることを意味する.

IS は静的に積分を求めるものであるが, これを時系列データ $\mathbf{x}_1 \cdots \mathbf{x}_T$ について拡張したものが Particle Filter (SMC とも呼ばれるが, 以下 PF) である.

紙面の都合上, 導出の詳細は割愛するが (導出については [8] がわかりやすい), PF では N 個のモンテカルロサンプル (パーティクルと呼ぶ) を準備し, その重み $w_t(\mathbf{x}^{(i)})$ $(i = 1 \ldots N)$ を $1/N$ で初期化し, 観測データ y_t がえられるごとに以下のように更新する.

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(y_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{q(\mathbf{x}_t|X_{t-1}, Y_t)} \tag{11}$$

$q(\mathbf{x}_t|X_{t-1}, Y_t)$ が提案分布であり，ここから $\mathbf{x}_t^{(1)} \dots \mathbf{x}_t^{(N)}$ をサンプルし, (11) 式に従って重み $w_t^{(i)}$ を更新する.

提案分布 q に制約がなく, 非線型な任意の分布を追跡することができるのが大きな特徴である. ここで, q が近似ではなく,

$$q(\mathbf{x}_t|X_{t-1}, Y_t) = p(\mathbf{x}_t|X_{t-1}, Y_t)$$

と解析的に正確に求まる場合には, (11) 式は簡単に次式となる.

$$w_t^{(i)} \propto w_{t-1}^{(i)} \cdot p(y_t|\mathbf{x}_{t-1}) \tag{12}$$

2.1 で述べたように, われわれの問題の場合, 変化点が与えられれば $p(\mathbf{x}_t|X_{t-1}, Y_t)$ はディリクレ分布として正確に求まることに注意されたい. このとき, PF による期待値は

$$E[y_t|y_1 \dots y_{t-1}] = \sum_{i=1}^{N} w_t^{(i)} E^{(i)}[y_t|y_1 \dots y_{t-1}] \tag{13}$$

である. すなわち, われわれの問題では, Particle Filter による事後分布は混合ディリクレ分布となる.

3.2 文脈の変化点検出問題

そこで次の問題は, 時間 t までの観測値 Y_t と, $(t-1)$ までの変化点系列 $\boldsymbol{I}_{t-1}$ が与えられたとき, 時間 t で変化が起こった確率 $p(I_t = 1|\boldsymbol{I}_{t-1}, Y_t)$ を求めることである.

ベイズの定理から,

$$p(I_t|\boldsymbol{I}_{t-1}, Y_t) \tag{14}$$
$$\propto p(I_t, y_t|\boldsymbol{I}_{t-1}, Y_{t-1}) \tag{15}$$
$$= p(y_t|Y_{t-1}, I_t, \boldsymbol{I}_{t-1})p(I_t|\boldsymbol{I}_{t-1}) \tag{16}$$
$$= \begin{cases} p(y_t|Y_{t-1}, \boldsymbol{I}_{t-1}, I_t = 1)p(I_t = 1|\boldsymbol{I}_{t-1}) & [\equiv a\,] \\ p(y_t|Y_{t-1}, \boldsymbol{I}_{t-1}, I_t = 0)p(I_t = 0|\boldsymbol{I}_{t-1}) & [\equiv b\,] \end{cases} \tag{17}$$

となるから, (17) 式をそれぞれ a, b とおけば,

$$p(I_t = 1|\boldsymbol{I}_{t-1}, Y_t) = \frac{a}{a+b} \tag{18}$$

$$p(I_t = 0|\boldsymbol{I}_{t-1}, Y_t) = \frac{b}{a+b} \tag{19}$$

と計算することができる.

(17) 式において, 第 1 項は変化/非変化が確定した後の出力 y の尤度であり, (3) 式および (6) 式から求まる. 第 2 項は変化の事前確率である. これは定数 ρ としてもよいが, PF においては各粒子が文脈の変化履歴 $\boldsymbol{I}_{t-1}$ を持つために, それを利用してオンラインで ρ の推定値を求めることができる.

すなわち, ρ がベータ事前分布 $\mathrm{Be}(\alpha, \beta)$ に従う確率変数であるとすると, $\boldsymbol{I}_{t-1}$ 中の 1 の回数を $n_{t-1}(1)$ とすれば, ベータ事後分布の期待値として,

$$E[\rho_t] = \frac{\alpha + n_{t-1}(1)}{\alpha + \beta + t - 1} \tag{20}$$

と ρ_t の推定値が求まる. 以下の実験では, すべてこのオンライン推定値を用いた.

次に, (12) 式における粒子の重みの更新係数 $p(y_t|\mathbf{x}_{t-1}) \equiv p(y_t|Y_{t-1}, \boldsymbol{I}_{t-1})$ について考えると,

$$p(y_t|Y_{t-1}, \boldsymbol{I}_{t-1}) \tag{21}$$
$$= \sum_{I_t \in \{0,1\}} p(y_t, I_t|Y_{t-1}, \boldsymbol{I}_{t-1}) \tag{22}$$
$$= \sum_{I_t \in \{0,1\}} p(y_t|I_t, \boldsymbol{I}_{t-1}, Y_{t-1})p(I_t|\boldsymbol{I}_{t-1}) \tag{23}$$
$$= p(y_t|I_t = 1, \boldsymbol{I}_{t-1}, Y_{t-1})p(I_t = 0|\boldsymbol{I}_{t-1})$$
$$+ p(y_t|I_t = 0, \boldsymbol{I}_{t-1}, Y_{t-1})p(I_t = 1|\boldsymbol{I}_{t-1}) \tag{24}$$
$$= a + b \tag{25}$$

と, (17) 式の a, b を用いて書けることがわかる.

以上により, Particle Filter により変化点を確率的に検出し, 予測を行うには,

1. 各粒子 $i = 1 \dots N$ について,
   (a) a, b を (17) 式に従って求める.
   (b) $I_t \sim \mathrm{Bernoulli}\,(a/(a+b))$ をサンプルして記録する.
   (c) 重み $w_t^{(i)} = w_{t-1}^{(i)} \cdot (a+b)$ と更新する.
2. $w_t^{(1)} \dots w_t^{(N)}$ および変化履歴 $\boldsymbol{I}_t$ から, (13) 式により予測確率を求める.

というアルゴリズムとなることがわかる.

現在の文脈からみて「変な」語 y_t が観測されると, 文脈予測確率 b よりもデフォルトの予測確率 a の方が高くなるため, 1(b) のベルヌーイ試行 $\mathrm{Bernoulli}\,(a/(a+b))$ において, 変化点 $I_t = 1$ がサンプルされやすくなる.

この変化点の検出は確率的に行うものであり, さらに N 個の粒子によって平均化されるために, 一回で文脈がすべて変わってしまう危険はないが, 続けてこれまでの文脈と違った語が現れた場合, そのどこかで文脈のシフトが起こることになる.

なお, 上記のステップ 1(c) において, 更新された重みに大きなばらつきが生じた場合, それに適応するために粒子を $w_t^{(i)} (i = 1 \dots N)$ に従って再サンプルし, 重みの小さな粒子を消し, 重みの大きいサンプルから「子供」を作る. この操作はリサンプリングとよばれているが, この際の基準として, 重みの変動係数 (CV) を用いるとよいことが知られている [9].

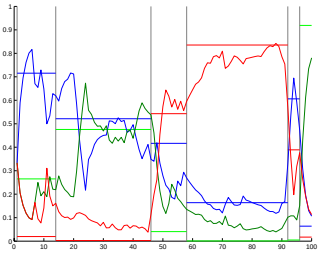


図 3: 図 1 に隠れた多項分布の, Particle Filter によるオンライン推定. 横線が真の θ_t である.

3.3 Multinomial Filtering

以上のアルゴリズムに従って, 図 1 の観測データから, 隠れた多項分布 θ_t を PF により推定したものが図 3 である. ただし, これは Forward 推定であり, 各 θ_t の推定において, 未来のデータは全く用いていないことに注意. ここでは粒子の数は $N = 50$, ベータ事前分布は $(\alpha, \beta) = (1, 10)$, CV の閾値 $= 1.0$ とした.

4 Mean shift model of Natural Language

Chen ら [6] はこの方法を, DNA 系列の推定に用いているが, これを自然言語の単語列にそのまま応用するにはいくつかの課題が残っている.

一つは, アルファベットの大きさが全く異なることである. ATGC の 4 種類しかアルファベットを持たない DNA と異なり, 自然言語には数万から数十万の単語が存在し, それらは独立ではなく, 互いに強い相関を持っている. たとえば, 「病院」という単語の後に「看護婦」という別の単語が多く出現しても, それらは関係が深く, 潜在的な変化は起こっていないと考えられるが, やはり別の記号である「大学」がその後に多く出現すれば, それは別の話題に移った (この場合, 「大学病院」というサブトピックに移った) と解すべきである. アルファベットを独立に扱う上記の MSM では, この関係はとらえることができない.

この関係をモデル化するために, テキストと単語の意味的な確率モデルである DM [22] と LDA [4] を用いて, MSM を自然言語に拡張した. この拡張により, 事前分布自体も [6] と異なり, 動的に更新することができる.

以下, DM と LDA について必要な解説を行いつつ, MSM-DM と MSM-LDA について述べる.

4.1 MSM-DM

Dirichlet Mixture (DM)[22] は, 文脈推定のために山本らによって近年提案された, 確率的なテキストモデルである. DM では, テキストのもつ多項分布の事前分布としてディリクレ分布ではなく, 混合ディリクレ分布を仮定し, その M 個の混合比 $\boldsymbol{\lambda} = \lambda_1 \dots \lambda_M$ と

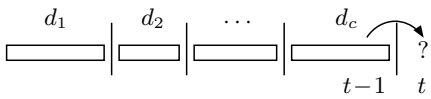


図 4: 変化点で仮想的な「文書」に区切られた履歴.

対応するディリクレ分布のハイパーパラメータ $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1 \dots \boldsymbol{\alpha}_M$ を, EM 法と Newton 法 (高速化のため, 実際には近似) を組み合わせることでコーパスから推定する.[4]

DM では, 履歴単語列 $\boldsymbol{h} = (w_1 w_2 \dots w_t)$ が与えられたとき, これを仮想的な (順序のない) 文書とみなし, 次式によって次の語 y を予測する. 詳細については, [22] を参照.

$$p(y|\boldsymbol{h}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) \propto \sum_{m=1}^{M} C_m \cdot \frac{\alpha_{my} + n(y)}{\alpha_m + h} \tag{26}$$

ここで $n(y)$ は $\boldsymbol{h}$ 中の y の生起回数, h は履歴の長さであり, C_m は次式である.

$$C_m = \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + h)} \prod_{v \in \boldsymbol{h}} \frac{\Gamma(\alpha_{mv} + n(v))}{\Gamma(\alpha_{mv})} \tag{27}$$

紙幅の都合で詳細は省略するが, [22] とは違った導出により, この方法は, 履歴から事前分布自体を適応的に選択し, C_m により適切に重みづけることで, 最適な予測を行うモデルであるとみなせる.

この DM を多項分布の MSM に用いるには, (3) 式および (6) 式において, y の予測確率を DM のものと置き換えればよい. [6] では多項分布の事前分布にディリクレ分布を仮定しているのに対し, この方法は混合ディリクレ分布を用いることで, そのきわめて自然な拡張になっていることがわかる.

ただし, (3) 式において, y の予測値は履歴 Y_{t-1} に全く依存しないため, 変化点が起こった後の予測は DM の事前分布からとられることになり, 精度の悪化を招きやすい. いま, 1 つの粒子についてみると, これまでの変化点によって履歴は仮想的な「文書」に区切られており (図 4), この情報を用いて事前分布を更新できる.

すなわち, DM のパラメータ推定において, λ_m は

$$\lambda_m \propto \sum_i p_{im} \tag{28}$$

(p_{im} は文書 i が m 番目の事前分布から生まれた確率) として求めるが, この方法を動的に適用し, p_{im} を履歴中の仮想的な「文書」に対して計算して和をとることで, λ_m の事後分布を求めることができる. ここで (28) 式において, λ_m の事前分布が右辺の p_{im} の計算に間接的に含まれていることに注意.

[4] http://cl.naist.jp/~daiti-m/dist/dm/ でパッケージを公開している.

この計算のためには p_{im} $(i = 1 \ldots c)$ だけが必要なため, 履歴をすべて保存する必要はない. 変化点がサンプルされた時に, 最近の変化点からの p_{im} を新しく計算して追加し, 以後 p_{im} だけを保存すればよい. これはフィルタリングアルゴリズムとして重要な点である. [5]

4.2 MSM-LDA

これに対し, LDA を用いて拡張する MSM-LDA では, 単語の出現確率の多項分布ではなく, 潜在的なトピック空間の多項分布を追跡する.

Latent Dirichlet Allocation (LDA)[4] とは, Blei らによって提案されたテキスト集合の確率モデルであり, 潜在意味モデルとして知られる PLSI [13] のベイズ的な発展形である.

LDA はパラメータとして, M 個のトピックに関するディリクレ事前分布のパラメータ $\boldsymbol{\alpha}$ と, トピック毎のユニグラム確率 $\boldsymbol{\beta} = \{\, p(v|m) \,\}$ $(v = 1 \,..\, L, m = 1 \,..\, M)$ をもつ. [6]

履歴 $\boldsymbol{h}$ が与えられたとき, LDA を用いた文脈モデル [21] では, 同様に $\boldsymbol{h}$ を仮想的な文書とみなし, 次の変分ベイズ EM アルゴリズムによって履歴のもつ潜在的なトピック分布 $q(\boldsymbol{\lambda}|\boldsymbol{h})$ を求める.

VB-E step:

$$q(z_i^t = 1|\boldsymbol{h}) \propto\ p(w_i|t) \exp(\Psi(\alpha + n_t)) \tag{29}$$

VB-M step:

$$q(\boldsymbol{\lambda}|\boldsymbol{h}) \propto\ \prod_{t=1}^{K} \lambda_t^{\alpha+n_t-1} \tag{30}$$

$$n_t = \sum_{i=1}^{h} q(z_i^t = 1|\boldsymbol{h}) \tag{31}$$

$q(\boldsymbol{\lambda}|\boldsymbol{h})$ はトピックの M 次元空間におけるディリクレ分布であり, トピックから単語への写像 $\boldsymbol{\beta}$ を用いて, 下のように次の語を予測する.

$$p(y|\boldsymbol{h}) = \int p(y|\boldsymbol{\lambda}) q(\boldsymbol{\lambda}|\boldsymbol{h}) d\boldsymbol{\lambda} \tag{32}$$

$$= \sum_{m=1}^{M} p(y|m) E_q[\lambda_m|\boldsymbol{h}]. \tag{33}$$

LDA を用いた MSM では, 単語の出現確率 $\boldsymbol{p}$ ではなく, 潜在的なトピック分布 $\boldsymbol{\lambda}$ を履歴から求めて追跡する. 具体的には, (2) 式と (4) 式において, 予測分布 $p(\theta_t|y_c \cdots y_{t-1})$ がトピック分布 $q(\boldsymbol{\lambda}_t|y_c \cdots y_{t-1})$ になる. 各粒子について, 上記変分ベイズ法により, 履歴から $q(\boldsymbol{\lambda}_t|y_c \cdots y_{t-1})$ を求め, (33) 式による語の予測を粒子全体について混合し, 最終的な予測を得る. 粒子の持つ各トピック分布はディリクレ分布であるから, この場合もトピックの事後分布は混合ディリクレ分布となる.

MSM-LDA においても, (30) 式の事前分布パラメータ α を履歴から更新できる. すなわち, 図 4 のように仮想的に「文書」に区切られた履歴において, 各「文書」$d_1 \,..\, d_c$ にはトピック事後分布 $q(\boldsymbol{\lambda}|d_i)$ $(i = 1 \,..\, c)$ が存在し, これらに共通するディリクレ事前分布を線形オーダーの Newton 法により求めることができる. 詳細については [4] 参照. 変化点がサンプルされるごとにこの計算を行うことで, 各粒子の持つ事前分布を更新することができる.

最初の事前パラメータ $\boldsymbol{\alpha}$ は Newton 法では使われないが, $q(\boldsymbol{\lambda}|d)$ を求める際に間接的に使われていることに注意. この Newton 法の計算にも全ての履歴を保存する必要はなく, 変化点ごとに $q(\boldsymbol{\lambda}|d)$ を計算し, 保存しておけばよく, オンラインアルゴリズムとなる.

5 実験と考察

British National Corpus (BNC) [5] を使って実験を行った. BNC はトピックが限定される WSJ 等と異なり, 様々なトピックが含まれるバランスドコーパスであり, このような実験に適している.

実験には BNC の Written テキスト 3,043 ファイルのうち, ランダムに選んだ 100 ファイルを評価データ, 残りを LDA/DM のパラメータ推定のための訓練データとした.

5.1 訓練データ

ただし, BNC のテキストは非常に長く (平均約 55,000 語), そのままの長さでは LDA および DM のパラメータを求めることができない.[7] 提案手法は一文書に関するモデルであるものの, 原理的には文書集合にも対しても拡張可能と考えられるが[8], 本稿の範囲を超えるため, ここでは近似として, 予備実験により, モデルの性能が低下しない最小のユニットとして 10 文[9] を採用し, 訓練セットの各テキストを 10 文毎に分割して文書としたものを訓練文書群とした.

ただし, BNC のデータは膨大であるため, 計算量の問題から, 訓練データのそれぞれのファイルを上記に従って分割し, 1 ファイルあたり最大 20 文書をランダムに抽出したものを最終的な訓練データとした. 最終的に, LDA/DM のパラメータ推定のための文書数は 56,939 文書, 11,032,233 語のデータとなっ

[5] さらに各 p_{im} は条件付き独立なため, 必要に応じて古い p_{im} を破棄しても, 他には影響を及ぼさない.

[6] `http://cl.naist.jp/~daiti-m/dist/lda/` でパッケージを公開している.

[7] 実際には, 50 文以上を 1 つのテキストとした場合にモデルが収束しなかった. これはテキストを 1 つの BOW とみなすテキストモデルが, 通常みられる, ある程度長い文書の集合に対しては無力という限界をもつことを示している.

[8] この場合, ベータ分布のハイパーパラメータ (α, β) を経験ベイズ法により推定できると考えられる.

[9] 以下, `<s>..</s>` で区切られる BNC のセグメント (ほぼ 1 文に対応する) を「文」と呼ぶ.

表 1: LDA/DM 訓練データの詳細

BNC 文書ファイル	2,943 ファイル
文書分割単位	$10 \leq \|d\| < 20$ 文
文書総数	56,939 文書
総語数	11,032,233 語
語彙数	52,846 語 (頻度 ≥ 5)

表 2: 評価用テキストの性質

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

た．これは BNC 全体の約 1/10 に相当する．語彙は頻度 5 以上の 52,846 語である．以上のデータを表 1 にまとめる．

5.2 評価データ

提案手法は, 文書内の文脈の動的な変化をとらえるモデルであり, 変化の速度自体も事後分布としてオンラインで求めつつ, 予測語の推定を行うものである．

この評価のためには, 様々な速度で変化するテキストが必要となるが, ここでは [20] にならい, 長いテキストから間隔を変化させてサンプリングを行うことで 4 種類の評価テキストを作成した．手順は [20] とほぼ同様であり, 以下のように行う．

(1) 各テキストに対し, 最初の文をランダムに選ぶ.
(2) その文から, 連続する X 文を採取する.
(3) Y 文だけスキップする.
(4) 求める文数のテキストが得られるまで, (2)(3) を繰り返す.

上記手順において, X, Y は表 2 に従う乱数である．この手順にしたがい, 種類毎に評価セットの各文書について 100 文をサンプルし, 評価用テキストとした．

5.3 実験設定

LDA および DM のパラメータ推定においては, それぞれクラス数を DM=50, LDA=200 とした．これは, 現在の Dirichlet Mixture の実装がハイパーパラメータに関して最尤推定になっているため, 混合数が少ない方が高い性能を持つからである [22].

文脈変化率を表すベータ分布の事前パラメータは, 原理的には一様分布 $(\alpha, \beta) = (1, 1)$ としてよいが, ここでは予備実験の結果から, $(\alpha, \beta) = (1, 50)$ とした．

5.4 実験結果

表 3 に, 各評価テキストセットに対する MSM-LDA, MSM-DM, LDA, DM のユニグラムパープレキシティを示す.

表 3: 各テキストセットに対するパープレキシティ

Text	MSM –DM	DM	MSM –LDA	LDA
Raw	**870.06**	925.83	**1028.04**	1037.42
Slow	**893.06**	974.04	**1047.08**	1060.56
Fast	**898.34**	988.26	**1044.56**	1061.01
VFast	**960.26**	1038.89	1065.15	1050.83

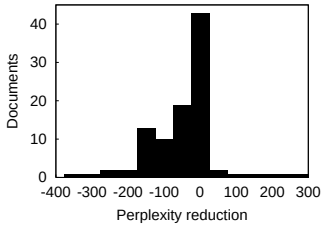


図 5: Dirichlet Mixture に対する, 評価データの各文書のパープレキシティ減少 ($\mathrm{PPL}_{MSM} - \mathrm{PPL}_{DM}$).

MSM-LDA においては精度上昇はわずかであるが, MSM-DM においては常にパープレキシティが減少しており, 文脈長を適応的に選択する効果があることがわかる.

図 5 に MSM-DM の, 'Raw' セットの各文書に対するパープレキシティ減少のプロットを示す．ほとんどの文書で効果があり, DM に比較して最大 400 程度パープレキシティが減少していることがわかる.

ただし実際には, 単語ごとに変化点をサンプルしているために, 提案法はノイズに比較的弱く, 時によって単語のパープレキシティが著しく (1000 倍程度) 増加する場合がある．これが図 5 にみえるほどの全体の精度上昇を生まない原因となっている.

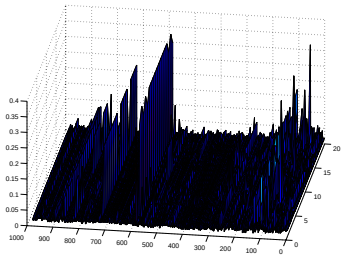
この問題を解決するためには, 変化点を単語ごとではなく, 文ごとなどに取ることが考えられるが[10] , テキストの生成モデルとしての単位は単語単位であり, PF において複数の観測値をまとめて扱うことのできる方法は見つかっていない. [16]

最後に, 評価テキストの一つの最初の 1000 語に対する, MSM-DM の文脈変化確率 $\boldsymbol{I}_t$ のプロットを図に示す．横軸が時間, 縦軸が粒子である．これからわかるように, 本手法は補助的に, TextTiling[12] の確率化を行うものともとらえることができる.

6 まとめ

本論文では, Mean Shift Model を DM および LDA によって拡張し, 文脈の変化点を動的にとらえる言語モデルを提案した．各粒子によってサンプルされた様々な長さの履歴からの予測を混合することで, 文脈をとらえた安定した予測が行われる．提案モデルは

[10] 単純に文の各語の確率の積を用いると, 式 (17) において二つの確率の差がきわめて大きくなってしまい, 変化点として 0 または 1 がほぼ確定的にサンプルされてしまう.



Forward モデルであり, これを Forward-Backward および文書集合へ適用することは今後の課題である.

謝辞: 本研究は独立行政法人 情報通信研究機構の研究委託により実施したものである。

参考文献

[1] Doug Beeferman, Adam Berger, and John Lafferty. A Model of Lexical Attraction and Repulsion. In *Proc. of ACL-EACL '97*, pages 373–380, 1997.

[2] Jerome R. Bellegarda. A Multispan Language Modeling Framework for Large Vocabulary Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 6(5):468–475, 1998.

[3] David Blei and Pedro Moreno. Topic Segmentation with an Aspect Hidden Markov Model. In *Proc. of SIGIR 2001*, pages 343–348. ACM Press, 2001.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Gavin Burnage and Dominic Dunlop. Encoding the British National Corpus. *English Language Corpora: Design, Analysis and Exploitation*, pages 79–95, 1992.

[6] Yuguo Chen and Tze Leung Lai. Sequential Monte Carlo Methods for Filtering and Smoothing in Hidden Markov Models. Discussion Paper 03-19, Institute of Statistics and Decision Sciences, Duke University, 2003.

[7] H. Chernoff and S. Zacks. Estimating the Current Mean of a Normal Distribution Which is Subject to Changes in Time. *Annals of Mathematical Statistics*, 35:999–1018, 1964.

[8] Arnaud Doucet. On Sequential Simulation-Based Methods for Bayesian Filtering. Technical Report CUED/F-INFENG/TR 310, Department of Engineering, Cambridge University, 1998.

[9] Arnaud Doucet, Nando de Freitas, and Neil Gordon. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer-Verlag, 2001.

[10] Daniel Gildea and Thomas Hofmann. Topic-based Language Models Using EM. In *Proc. of EUROSPEECH '99*, pages 2167–2170, 1999.

[11] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall / CRC, 1996.

[12] Marti Hearst. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9–16, 1994.

[13] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. of SIGIR '99*, pages 50–57, 1999.

[14] Frederick Jelinek. *Statistical Methods for Speech Recognition*. Language, Speech, and Communication Series. MIT Press, 1998.

[15] Sadao Kurohashi and Manabu Ori. Nonlocal Language Modeling based on Co-occurence Vectors. In *Proc. of EMNLP/VLC '00*, pages 80–86, 2000.

[16] Cody Kwok, Dieter Fox, and Marina Meilă. Real-time Particle Filters. In *Advances in Neural Information Processing Systems 15*, 2002.

[17] Peter M. Lee. *Bayesian Statistics: An Introduction*. Arnold Publishers, Second edition, 1997.

[18] Philippe Sollers. *H*. Seuil (1 mars 1973) edition, 1973.

[19] Yi-Chin Yao. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *Annals of Statistics*, 12:1434–1447, 1984.

[20] 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*, pages 107–112, 2003.

[21] 三品 拓也, 山本 幹雄. 確率的 LSA に基づく ngram モデルの変分ベイズ学習を利用した文脈適応化. **信学技報** *NLC2002-73*, pages 13–18, 2002.

[22] 山本 幹雄, 貞光 九月, 三品 拓也. 混合ディリクレ分布を用いた文脈のモデル化と言語モデルへの応用. **情報処理学会研究報告** *2003-SLP-48*, pages 29–34, 2003.

[23] 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** *2004-SLP-53*, pages 1–6, 2004.