Gibbs Sampling による確率的テキスト分割と複数観測への拡張

持橋大地　菊井玄一郎
ATR 音声言語コミュニケーション研究所
daichi.mochihashi@atr.jp, genichiro.kikui@atr.jp

1　はじめに

文書分類や情報検索などで用いられる文書モデルでは多くの場合, bag-of-words 表現が使われ, 単語の時系列情報が落ちていることが多いが, 実際には文書の内部は均質ではなく, 様々な話題が順に呈示されることで構造化されている.

このような文書内部の意味的遷移をモデル化するために, われわれは文書内部に隠れた意味的変化点があると仮定し, 確率的テキストモデルである LDA および DM[4] と組み合わせることで, 逐次モンテカルロ法を用いて変化点をオンライン推定し文脈を追跡する方法を示した [1][2].

しかしながら, この方法は隠れた変化点を単語ごとに計算するためにノイズに弱く, 現在の文脈にあまり関係ない語が偶然現れた場合にそこを変化点として検出してしまい, 時として予測が著しく悪化するという欠点があった. 図 1 に, ある文書 (1,910 語) における提案法と, 文脈として過去の履歴をすべて用いる従来法との予測確率の比を, 文書の先頭からの位置を横軸にとって示す. 全体として予測が良くはなっているものの (y 軸 < 1), 縦軸の対数比でみると予測確率がかなり悪くなってしまう場合もあることがわかる.
もう一つの欠点として, 提案した方法は予測モデルであるため, 変化点確率の計算は最近の観測語および過去の履歴のみに基づいて行われている (いわゆる Forward モデル) という点が挙げられる. 音声認識等における応用ではこの方法が必要であるが, 文書全体が先に与えられており, それをモデル化する際には, 後ろの情報も用いることで文脈の変化をより正確にとらえ, 変化点確率を計算することができると考えられる (いわゆる Forward-Backward モデル).

そこで本稿では, まず変化点を複数の観測値を持つブロック毎に計算し, 推定をロバスト化することを試み, その結果を報告する. ただし, ブロック長があまり大きくなるとテキストの潜在的なダイナミクスが捉えられず, 推定がかえって悪くなることが予想される. 実際に, 数単語程度のブロックに取ることが予測を最良にすることがわかった.

提案法は非線形な HMM であるため, Forward-Backward 計算には通常の Baum-Welch アルゴリズムを適用することはできない. このため, 変化点を次々とサンプリングしていく Gibbs Sampler による解法を示し, 解析例を考察する. 最後に TextTiling 法 [7] との比較など今後の課題について述べ, 文書に対する言語モデルとしてソフト分かち書き [5] との関連についてふれる.

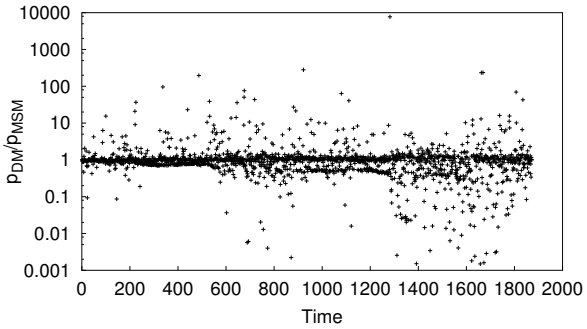


図 1: DM/MSM-DM の予測確率比と文脈長.

2　MSM-DM と意味的変化点確率

変化点を持つ確率的テキストモデルとして提案したものに, LDA を用いた MSM-LDA と DM を用いた MSM-DM を提案した [1, 2] が, 後者の方が単語単体をすべてモデル化できるため柔軟であり, 計算も高速で文脈モデルとして高性能であることがわかっているため, 以下では MSM-DM について考える.

2.1　MSM-DM の予測モデル

逐次モンテカルロ法 (SMC) を用いた MSM-DM の解法では, 各モンテカルロサンプル (粒子とよばれる) が変化点を実際に 0/1 でサンプリングし, 粒子ごとに異なる最近の変化点からの履歴を用いて予測を行う. この予測分布を, 各粒子について適切に更新される重みに基づいて混合することで最適な予測分布を生成する, というアルゴリズムになっている. 図 2 に, この予測アルゴリズムの動作を示す.

2.2　変化点確率の拡張

上記の SMC における解法では, 各粒子は時間 t での変化点 $I_t = \{0, 1\}$ を, 次の二項確率に従ってサンプリングする. ここで, $\mathbf{w}_t$ は観測値 (単語または単語ブロック) $w_1 w_2 \cdots w_t$, $\mathbf{I}_t$ は過去の I_t の履歴 $\mathbf{I}_t = \{I_1, I_2, \ldots, I_t\}$ であるとする.

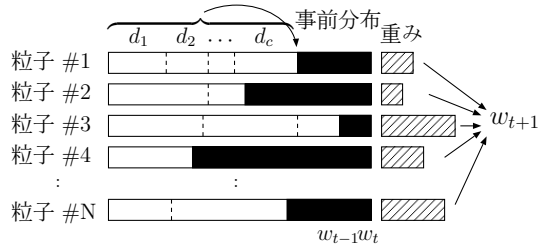


図 2: MSM-DM の予測モデル.

$$p(I_t|\mathbf{I}_{t-1},\mathbf{w}_t) \propto p(w_t|\mathbf{w}_{t-1},\mathbf{I}_{t-1},I_t)p(I_t|\mathbf{I}_{t-1}) \quad (1)$$

$$= \begin{cases} p(w_t|\mathbf{w}_{t-1},\mathbf{I}_{t-1},I_t=1)p(I_t=1|\mathbf{I}_{t-1}) = f(t) \\ p(w_t|\mathbf{w}_{t-1},\mathbf{I}_{t-1},I_t=0)p(I_t=0|\mathbf{I}_{t-1}) = g(t) \end{cases}$$

w_t のそれぞれが複数の観測値 (ブロック長 l) からなるとき, この確率は以下のように展開できる.

Case $I_t = 1$:

$$f(t) = \rho \cdot \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m+l)} \prod_{v\in w_t} \frac{\Gamma(\alpha_{mv}+n_v(w_t))}{\Gamma(\alpha_{mv})} \quad (2)$$

Case $I_t = 0$:

$$g(t) = (1-\rho) \cdot \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m+h)}{\Gamma(\alpha_m+h+l)} \prod_{v\in w_t} \frac{\Gamma(\alpha_{mv}+n_v(\mathbf{h})+n_v(w_t))}{\Gamma(\alpha_{mv}+n_v(\mathbf{h}))} \quad (3)$$

ここで $\mathbf{h}$ は各粒子ごとに異なる, $\mathbf{w}_{t-1}$ の中で最近の変化点以後の観測値, h はその長さであり, $n_v(x)$ は x 中の v の出現回数を表す.

Γ 関数の性質 $\Gamma(x+1)=x\Gamma(x)$ から, $l=1$ のとき上式は [1, 2] における式と一致し, その自然な拡張となっている.

この確率に従い, 各粒子はベルヌーイ試行

$$I_t \sim \text{Bernoulli}\left(\frac{f(t)}{f(t)+g(t)}\right) \quad (4)$$

を行って変化点をサンプルし, 次の観測を予測, 粒子の重みを更新する. 変化点の事前確率 ρ は $\mathbf{I}_{t-1}$ から, ベータ事後分布の期待値

$$\langle \rho_t \rangle = \frac{\alpha + n_{t-1}(1)}{\alpha+\beta+t-1} \quad (5)$$

として求まる. ここで $n_{t-1}(1)$ は $\mathbf{I}_{t-1}$ 中の 1 の回数. この事前分布として何が適切かはブロック長 l によって異なるため, 実験ではすべてハイパーパラメータ $(\alpha,\beta)=(1,1)$ とし, 一様事前分布を用いた.

3 意味的変化点の Gibbs Sampling

文脈から次の語を次々に予測し, 実際に観測された語に従って予測分布を更新していく SMC アルゴリズムの他に, 文書全体が与えられた下での変化点をサンプリングしていく Gibbs Sampler を構成することができる.

1 $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$ をランダムに 0/1 で初期化. (Bernoulli(p) からの乱数)
2 For $t = 1..T$ {
  For n = randperm(N) {
    Sample $I_n^{(t)} \sim p(I_n|\mathbf{I}^{(t)}\backslash n, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$.
  }
}

図 3: Gibbs Sampler of MSM-DM.

3.1 Gibbs Sampler of MSM-DM

ギブスサンプリングとは, マルコフ連鎖モンテカルロ法 (MCMC) の単純な場合であり, 隠れ変数を一つずつサンプリングして更新してゆくことで, 真の隠れ変数の分布に従ったサンプルを得る方法である [6]. われわれの場合, テキスト $\mathbf{w} = w_1 w_2 \cdots w_N$ に対して, 隠れた変化点ベクトル $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$ をサンプリングする, 図 3 のアルゴリズムとなる.

ここで randperm(N) は $1 \ldots N$ のランダムな並び換えを返す関数であり, $\mathbf{I}^{(t)}\backslash n$ は $\mathbf{I}^{(t)}$ から $I_n^{(t)}$ を除いた集合. 初期化にかかわらず, 繰り返し回数 $T \to \infty$ の極限で, 図 3 のアルゴリズムから得られる変化点ベクトル $\mathbf{I}^{(t)}$ は隠れた真の分布からのサンプルに収束し,

$$I_n = \frac{1}{T}\sum_{t=1}^{T} I_n^{(t)} \quad (6)$$

とすることで時刻 n での変化点確率 $p(I_n=1)$ が計算できる.[1]

3.2 条件つき確率の計算

図 3 のアルゴリズムの中で, 変化点を実際にサンプルする確率分布 $p(I_n|\mathbf{I}\backslash n, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ は以下のように変形できる.

$$p(I_n|\mathbf{I}\backslash n, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \propto p(I_n, \mathbf{I}\backslash n, \mathbf{w}|\boldsymbol{\lambda}, \boldsymbol{\alpha}) \quad (7)$$

$$\propto p(\mathbf{w}|I_n, \mathbf{I}\backslash n, \boldsymbol{\lambda}, \boldsymbol{\alpha})\, p(I_n, \mathbf{I}\backslash n|\boldsymbol{\lambda}, \boldsymbol{\alpha}) \quad (8)$$

$$\propto p(\mathbf{w}|I_n, \mathbf{I}\backslash n, \boldsymbol{\lambda}, \boldsymbol{\alpha})\, p(I_n|\mathbf{I}\backslash n). \quad (9)$$

(9) 式において第 2 項は (5) と同様にして求まる. 第 1 項は n より前の最近の変化点を B, 後の最近の変化点を F として,

$$\begin{cases} p(\mathbf{w}|I_n=1, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \propto p(w_B \ldots w_{n-1}|\boldsymbol{\lambda}, \boldsymbol{\alpha}) \\ \qquad\qquad \times p(w_n \ldots w_F|\boldsymbol{\lambda}, \boldsymbol{\alpha}) \\ p(\mathbf{w}|I_n=0, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \propto p(w_B \ldots w_F|\boldsymbol{\lambda}, \boldsymbol{\alpha}) \end{cases} \quad (10)$$

によって DM の文書確率から求めることができる. すなわち, (9)(10) 式は変化点 B と F で挟まれたテキストが, n を挟んで

[1] 実際には, 小さい t に対する $\mathbf{I}^{(t)}$ は初期値に依存するため, 最初の方の値 (burn-in) は捨てて計算することで, 精度のよいサンプルが得られる.

- $w_B \cdots w_{n-1}, w_n \cdots w_F$ が別に生成された確率 (変化点あり)
- $w_B \cdots w_F$ が同時に生成された確率 (変化点なし)

を変化点の事前確率で重みづけて比べ, 前者の確率が高ければ時間 n での変化点確率が高まる, という式になっている. ここで, 同じ内容語が前と後に同時に出現していた場合, DM の下ではこの場所の変化点確率は非常に低くなることに注意したい. DM は混合 Polya 分布を持つが, Polya 分布の下では同じ語の 2 回以上の出現回数がダンピングされ, 独立な場合より高い確率を持つようになるため [8], 前後を分割しない方がずっと高い確率を与えるからである.

なお, この方法は, Steyvers and Brown [3] の方法と ρ を動的に更新する部分を除いて同じである. [3] は認知過程のモデルとして, 比較的低次元な二項分布が用いられているが, 提案法は DM を用いて, それを意味的相関を持つ数万次元の記号列に拡張したものになっている.

3.3 周辺化パープレキシティと変化点

このアルゴリズムは $p(\mathbf{I}|\mathbf{w})$ に従って変化点系列 $\mathbf{I}$ をサンプルする方法であるが, $p(\mathbf{I}|\mathbf{w}) \propto p(\mathbf{I}, \mathbf{w})$ であるから, この方法で変化点を求めることは, 文書と変化点の結合確率 $p(\mathbf{w}, \mathbf{I})$ において, 変化点 $\mathbf{I}$ を周辺化した文書確率

$$p(\mathbf{w}) = \sum_{\mathbf{I}} p(\mathbf{w}, \mathbf{I}) \tag{11}$$

の和に含まれる各サンプルを得ていることに相当している. $\mathbf{I}$ の可能な組み合わせは 2^N 個と厖大になるが, ギブスサンプリングによりその中で確率の高いものを確率的に取り出し, (11) 式を以下で近似することができる.

$$p(\mathbf{w}) \simeq \frac{1}{T} \sum_{t=1}^{T} p(\mathbf{w}, \mathbf{I}^{(t)}) \tag{12}$$

文書 $\mathbf{w}$ のパープレキシティは $p(\mathbf{w})^{-1/N}$ であるから, 各変化点サンプルに対して, そこで $\mathbf{w}$ が区切られた文書確率を計算し幾何平均を取ることで, 隠れた確率的な変化点を考慮した "周辺化パープレキシティ" を求めることができる.

4 実験および考察

4.1 実験設定

実験には 2001 年度版毎日新聞記事のうち, テストデータとして, 1,000 語以上の記事 3,250 記事からランダムに選択した 100 記事 (平均 1454.4 語) を用い, 1,000 語を超えない 97,131 記事, 22,912,999 語を Smoothed DM のパラメータ推定のための訓練データとした. 語彙は頻度 5 以上の 65,313 語である. DM の混合数は 200 とした.[2]
MeCab-0.8.1 を用いて分かち書きを行い, 文についての実験では句点を文の境界とした.[3]

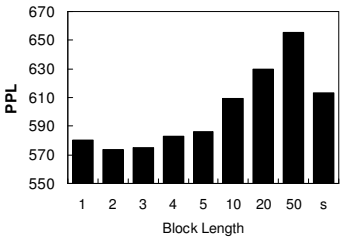
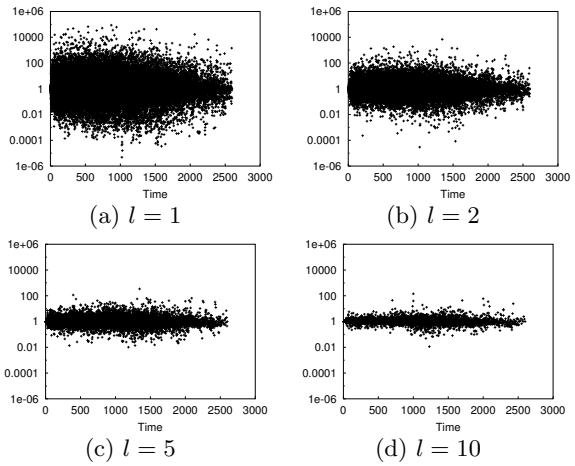


図 4: ブロック長 l での予測パープレキシティ. "s" は文単位の変化点を表す.

4.2 複数観測での MSM-DM

図 4 に, ブロック長 $l = 1 \sim 5, 10, 20, 50$ および文単位とした各場合におけるテストセットの予測パープレキシティを示す. $l = 1$ が単語ベースのモデルである. 文を単位とすることが特別良いということはなく, むしろ数語を単位としてダイナミクスを柔軟にモデル化した方が, 全体として高い予測確率を与えることがわかる.

図 5 に, 文脈長を横軸に, 単純な DM との予測確率の比 $p_{\mathrm{DM}}/p_{\mathrm{MSM}}$ を縦軸に取ったプロットを示す. ブロック長 l が増えるに従い, 予測が大きく悪化する (y 軸 $\gg 1$) ことは少なくなるものの, 単純な DM の予測に近づいてしまい, 大きく改善する ($\ll 1$) ことも少なくなる, というトレードオフがあることがわかる.



(a) $l = 1$ (b) $l = 2$

(c) $l = 5$ (d) $l = 10$

図 5: 各ブロック長 l における予測確率比.

[2] Smoothed DM [4] のパッケージを `http://chasen.org/~daiti-m/dist/dm/` で公開している.

[3] これ以外の形態素情報は, 本論文全体を通じ全く用いていないことに注意.

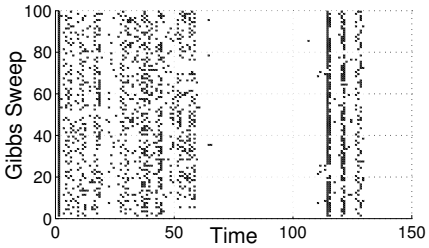


図 6: Gibbs sweep とテキストの変化点.

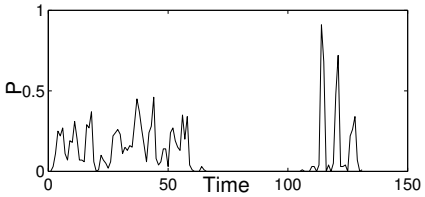


図 7: Gibbs による変化点確率 (図 6 を正規化したもの).

4.3 変化点の Gibbs Sampling

図 6 に, 1,311 語のテキストに対する変化点のギブスサンプリングの例と, それを正規化することで (6) 式から得られる変化点確率を示す. ブロック長を 10 とし, $p = 0.5$ の確率で変化点をランダムに初期化, 100 回の burn-in の後に図 6 の 100 個のサンプルを得た. [4]

図 8 の SMC による前向き変化点確率と比べると, 前後の情報を用いることでノイズが減り, 推定がより正確になっていることがわかる.

この文書に対する DM の文書パープレキシティは 700.03, SMC による予測パープレキシティは 657.43, 周辺化パープレキシティは 644.91 であった.

5 まとめと今後の課題

先に提案した文脈の推定法をブロック単位の観測に拡張し, あわせて文書全体を考慮してテキストの意味的な変化点を Gibbs Sampling により求める方法を示した.

Gibbs Sampling により意味的変化点を求める手法は, TextTiling[7] のヒューリスティックなアルゴリズムを確率モデルの観点からとらえ直したようなものになっている. しかしながら, ブロック間の cosine 距離に何ら明確な意味がなく, モデル的な裏付けのない TextTiling と異なり, 提案手法は生成モデルから導かれる変化点確率として明確な意味を持っており, 閾値を用いて変化点を決定することなしに, 確率のまま 'ソフトな' 取り扱いが可能になるという利点を持っている. 今後, 比較実験を行うことでこの利点を検証していきたい.

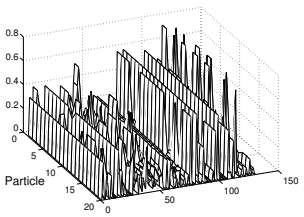


図 8: 図 7 のテキストの, SMC による変化点確率.

工藤 [5] は, マルコフ確率場に逆温度パラメータ θ を導入し動的計画法を用いることで, 形態素解析における区切りを同様に確率化できることを示した. これは形態素が確率的に区切られる言語モデルにつながると考えられるが, その際, 本稿で提案したようにパラメータ自体が言語ストリームの中で変化し, 形態素解析結果が文脈によって異なるモデルが考えられる. 意味的に確率的に区切られ, さらに形態素も確率的に区切られ相互作用を持つような, そんな新しい言語モデルを構築していきたい.

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「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] 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 自然言語処理研究会** *2005-NL-165*, 2005.

[2] Daichi Mochihashi and Yuji Matsumoto. Context as Filtering. In *Advances in Neural Information Processing Systems 18*, 2005.

[3] Mark Steyvers and Scott Brown. Prediction and Change Detection. In *Advances in Neural Information Processing Systems 18*, 2005.

[4] 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.

[5] 工藤拓. 形態素周辺確率を用いた分かち書きの一般化とその応用. In **言語処理学会全国大会論文集** *NLP-2005*, 2005.

[6] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, 2nd Edition*. Chapman & Hall/CRC, 2003.

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

[8] Thomas P. Minka. Estimating a Dirichlet distribution, 2000. http://www.stat.cmu.edu/~minka/papers/dirichlet/.

[4] MATLAB による実装では, このサンプリングには 7 時間 24 分を要した. (Xeon 2.8GHz)