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

daichi.mochihashi@atr.jp

ATR 音声言語コミュニケーション研究所

言語処理学会 年次大会 2006 Poster P2–5

概要 (1)

- 文脈の変化をとらえる, Particle Filter による推定法
  - 意味的変化点の確率的な計算
  - 単語ベースの変化点 · · · ノイズに弱い

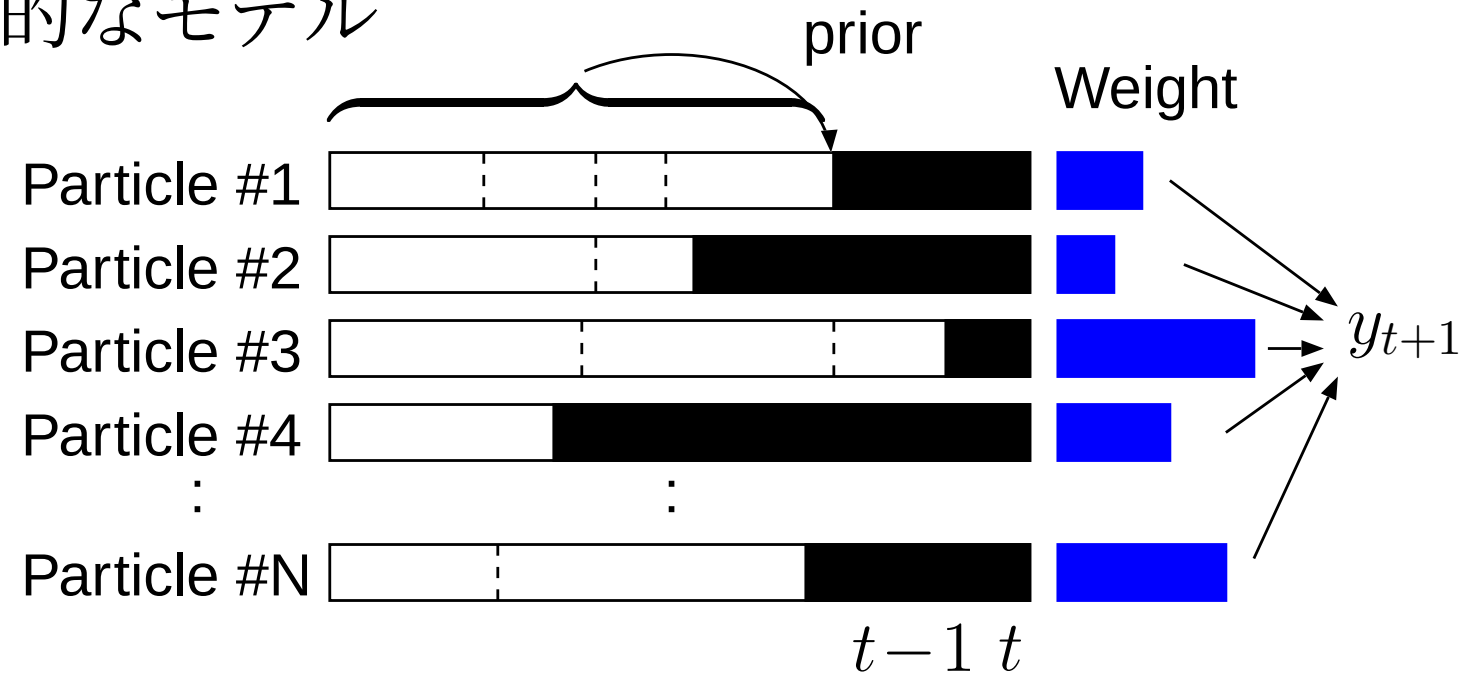
    ⇓
- 任意長の**ブロックベース**の変化点
  - ノイズに強い
  - ブロック長が長すぎると, テキストの潜在的なダイナミクスに追従できない · · · 最適なブロック長は?
  - 実験結果: 文ベースの予測は最適ではない

概要 (2)

- 文脈モデルの意味的変化点は, 前向き推定 (予測モデル)
  - 後の文章を見ていない!
- 文書全体を見渡した, 変化点確率の計算
  - 通常の Forward-Backward は使えない
    ⇓
  - **Gibbs Sampler** による推定
    - マルコフ連鎖モンテカルロ法 (MCMC) の単純な場合
    - 前後のブロックが意味的につながるか? を確率的にサンプリング.
- 文書内部の意味的変化点を考慮した,
  文書の**周辺化パープレキシティ**が計算可能

文脈の変化を SMC で追跡 / MSM-DM

- Dirichlet Mixtures (DM) (山本, 貞光 2004/2005) の動的なモデル



  - 文脈の意味的な変化点を, 並列にサンプリングしてシミュレーションする (SMC/Particle Filter)
  - 最後の変化点以後の観測値を文脈として用いて, 適切に混合
  - ある単語 y_t を見たとき, ここで文脈に変化が起こったか? を確率的にとらえる必要がある

文脈の変化を SMC で追跡 / MSM-DM (2)

- 時刻 t での変化点確率を以下で計算 ($\mathbf{h}$: 履歴)

$$p(I_t=1|\mathbf{h}) = \frac{f(t)}{f(t)+g(t)} \tag{1}$$

where

$$f(t) = p(I_t=1)p_{\mathrm{DM}}(y_t) \tag{2}$$

$$= \frac{\alpha + n_{t-1}(1)}{\alpha+\beta+t-1}\sum_m \lambda_m \frac{\alpha_{my}}{\sum_v \alpha_{mv}}, \tag{3}$$

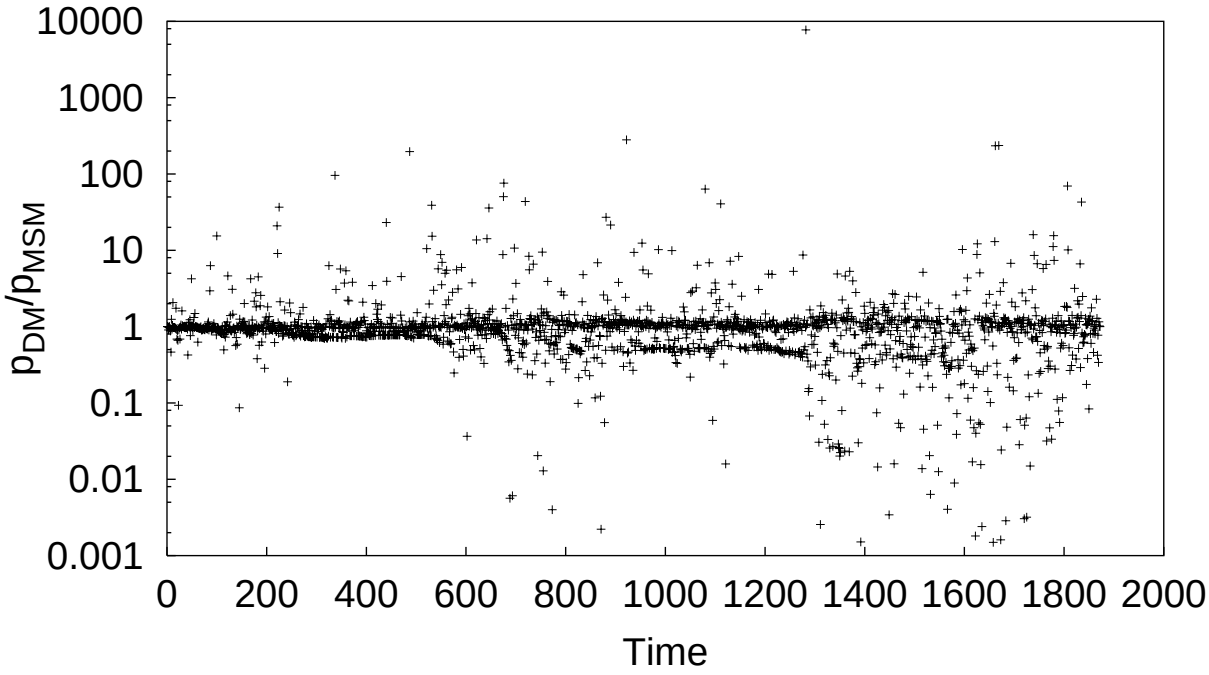
$$g(t) = p(I_t=0)p_{\mathrm{DM}}(y_t|\mathbf{h}) \tag{4}$$

$$= \left(1-\frac{\alpha + n_{t-1}(1)}{\alpha+\beta+t-1}\right)\sum_m C_m \frac{\alpha_{my}+\#(y\in\mathbf{h})}{\sum_v \alpha_{mv}+h}. \tag{5}$$

y_t : 時刻 t で観測した単語　h : $\mathbf{h}$ の長さ

文脈の動的な追跡による改善

- $p(\text{DM})/p(\text{MSM-DM})$ の予測確率比



  - y軸 < 1 が改善されている場合
  - どんな場合に改善されるか?
    - 時々, 非常に悪くなる時がある $\cdots$ ノイズの影響?

文脈の動的な追跡による改善 (2)

t	w	$p(\mathrm{DM})/p(\mathrm{MSM})$
501	熱心	0.4979
502	自分	0.7491
503	自身	0.4616
504	で	1.0433
505	も	1.1626
506	データ	0.4962
507	を	1.0140
508	取っ	0.7247
509	たり	0.4263
510	すごい	2.1357
:		

t	w	$p(\mathrm{DM})/p(\mathrm{MSM})$
1561	#	0.3416
1562	#	0.3392
1563	鈴木	0.5977
1564	誠	0.0529
1565	ロイヤルズ	3.9497e+03
1566	#	1.3864
1567	年	4.3377
1568	目	1.1804
1569	の	1.1263
1570	初	1.2883
:		

複数観測への拡張

- 各時刻 t で, 長さ l の単語ブロック $\mathbf{y}_t$ が観測されるとき,

$$f(t) = p(I_t = 1) p_{\mathrm{DM}}(\mathbf{y}_t) \tag{6}$$

$$= \rho \cdot \sum_m \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + h)} \prod_v \frac{\Gamma(\alpha_{mv} + \#(v \in \mathbf{y}_t))}{\Gamma(\alpha_{mv})} \tag{7}$$

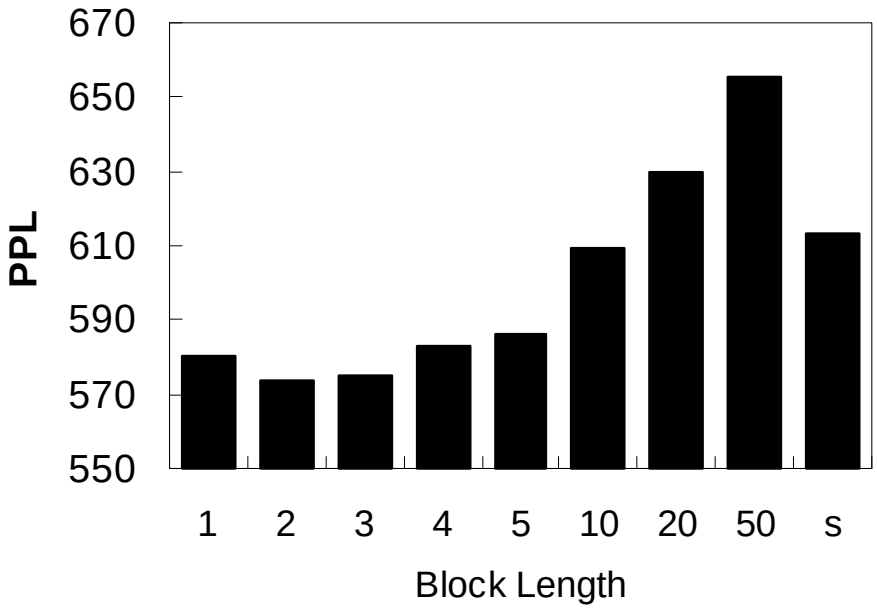
$$g(t) = p(I_t = 0) p_{\mathrm{DM}}(\mathbf{y}_t | \mathbf{h}) \tag{8}$$

$$= (1-\rho) \cdot \sum_m \lambda_m \frac{\Gamma(\alpha_m + h)}{\Gamma(\alpha_m + h + l)} \prod_v \frac{\Gamma(\alpha_{mv} + \#(v \in \mathbf{h}) + \#(v \in \mathbf{y}_t))}{\Gamma(\alpha_{mv} + \#(v \in \mathbf{h}))} \tag{9}$$

- $\Gamma(x+1) = x\Gamma(x)$ より, $l = 1$ のとき式 (3)(5) と一致

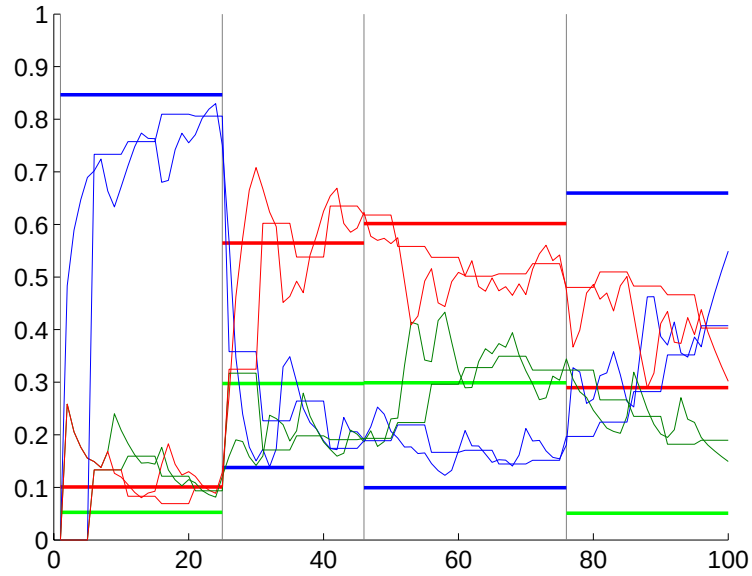
複数観測への拡張: 結果

- ブロック長 l での予測パープレキシティ



  - ブロック長が長すぎると, テキストの潜在的ダイナミクスに追従できない
  - 文レベルの予測は, 必ずしも最適ではない!

- Toy Model



  - 観測アルファベット:

    aaaaaaabaaaaaaacaaaaaaaac\
    cccccbcaacbccccabcccacbc\
    bbbccbbccacbbcbcbcacccbcb\
    aaccacaccbaaaccabcacaaaaa

  - ブロック長 $l = 1$ の場合と $l = 5$ の場合

文書全体を見て意味的変化点を推定 / Gibbs Sampler

- テキスト

  $\mathbf{w} = w_1 w_2 \cdots w_N$ の裏には, 隠れた意味的変化点

  $\mathbf{I} = I_1 I_2 \cdots I_N$ (たとえば, $10010 \cdots 000$) があると仮定

- テキスト $\mathbf{w}$ の確率 $p(\mathbf{w})$ は, $\mathbf{I}$ を周辺化することで

$$p(\mathbf{w}) = \sum_{\mathbf{I} \in \{0,1\}^N} p(\mathbf{w}, \mathbf{I}) \tag{10}$$

  で計算できる.

- $\mathbf{I}$ の可能な組み合わせは 2^N 個と, 天文学的な数字

  (マルコフ性を持たないため, Forward-Backward は使えない)

  $\rightarrow$ Gibbs Sampler により, $\mathbf{I}$ を確率的にサンプリング

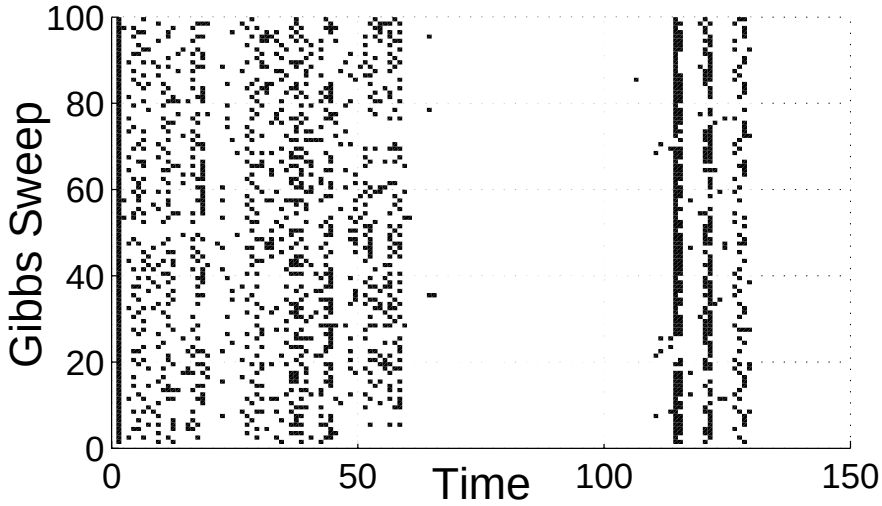
Gibbs Sampler (2)

- $\mathbf{I} = I_1 I_2 \cdots I_N$ に含まれる I_n を, 順番にアップデート
  1. $\mathbf{I}$ をランダムに 0/1 で初期化.
  2. For $t = 1 \ldots T$,
     For $n = \mathrm{randperm}(N)$,
     Draw $I_n^{(t)} \sim p(I_n | \mathbf{I} \backslash I_n, w_n, \boldsymbol{\lambda}, \boldsymbol{\alpha})$
  - 変化点系列のサンプル $\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \ldots, \mathbf{I}^{(T)}$ が得られる.
- $p(I_n | \mathbf{I} \backslash I_n, w_n, \boldsymbol{\lambda}, \boldsymbol{\alpha})$: 単語 w_n の場所での変化確率
  - 前後が繋がって生成された尤度 [変化点なし]
  - 前後が別々に生成された尤度 [変化点あり]

  を比べることで, 計算できる.

Gibbs Sampler (3)

- 1311 語のテキストに対する Gibbs sampler の結果



$\rightarrow \mathbf{I}^{(100)}$

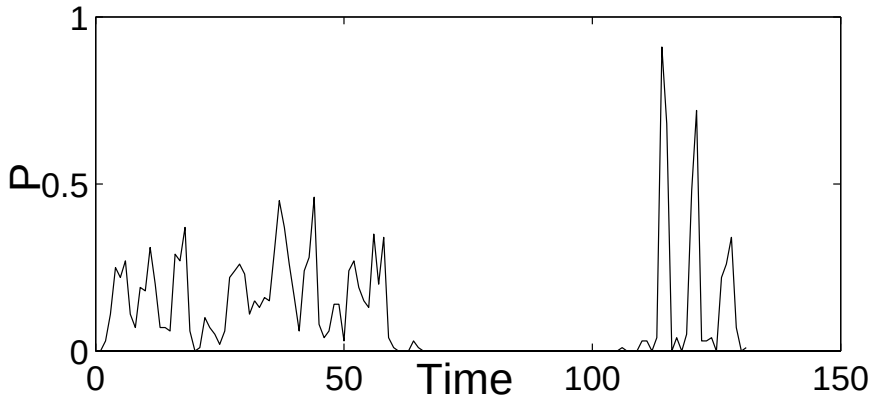
$\rightarrow \mathbf{I}^{(99)}$

$\vdots$

$\rightarrow \mathbf{I}^{(2)}$

$\rightarrow \mathbf{I}^{(1)}$

$\Downarrow$ 変化点確率



PPL = 700.03

Marginalized PPL = 644.91

SMC Predictive PPL = 657.43

周辺化言語モデル

- パープレキシティ

$$PPL(\mathbf{w}) = p(\mathbf{w})^{-1/N}. \tag{11}$$

- ここで, Gibbs sampling により,

$$p(\mathbf{w}) = \sum_{\mathbf{I}\in\{0,1\}^N} p(\mathbf{w},\mathbf{I}) \simeq \frac{1}{T}\sum_{t=1}^{T} p(\mathbf{w},\mathbf{I}^{(t)}) \quad \text{ゆえ,} \tag{12}$$

  - 周辺化パープレキシティ

$$PPL(\mathbf{w}) = \left(\frac{1}{T}\sum_{t=1}^{T} p(\mathbf{w},\mathbf{I}^{(t)})\right)^{-1/N}. \tag{13}$$

    - 隠れた意味的区切りを**確率的に**考慮
    - 閾値を用いる TextTiling のような方法では, 確率ではないために, このような計算は行えない.

まとめ

- 任意長 l のブロック毎の文脈推定モデル
  - $l = 2 \sim 3$ で平均予測確率最大
  - $l =$ 文単位 は必ずしも最適な予測とならない
    - DM がキャッシュモデルであるために, かえって予測が悪化する場合がまだ存在
- 文書全体に対して, Gibbs sampler を用いた意味的変化点推定
  - TextTiling と似ているが, 変化点が確率として求まる + 確率的トピックモデルとの連結
  - 変化点のすべての可能性を考慮した, 周辺化パープレキシティ
    - *cf.* 形態素の全ての区切り方を考慮した, 周辺化 MRF (ソフト分かち書き; 工藤 2005)
    - Semantics $\Leftrightarrow$ Syntax.

今後の課題

- TextTiling との比較実験
- $p(\mathrm{DM}) \gg p(\mathrm{MSM\text{-}DM})$ となる場合の修正
  - ベイズキャッシュモデルを超えるモデル化.
- 周辺化 MRF との Joint Modeling.
  (形態素モデル＋意味モデル)