

可変次数無限隠れマルコフモデル

内海 慶^{1,a)} 持橋大地^{2,b)}

概要：隠れマルコフモデルは系列データのモデルとして広く用いられる。自然言語処理では形態素解析や品詞推定のモデルとして、音声工学では音声認識で入力信号から音素を推定する音響モデルとしてよく利用される。従来、隠れマルコフモデルでは状態は1つ前の状態にのみ依存する、1次マルコフモデルが利用されてきた。この理由として、高次マルコフモデルでは計算量が次数に応じて指数的に大きくなること、及び状態遷移の組み合わせが膨大になることから、各遷移に対する学習事例が相対的に少なくなってしまう、データスパースネスの問題が起こることがあげられる。

本稿では、適切な事前分布を導入することでデータスパースネスの問題を解決し、また次数 n を確率変数として系列データの各位置に導入することで、隠れマルコフモデルを次数可変へ拡張した手法を提案する。提案手法によって、高次隠れマルコフモデルを効率的に計算し、また従来では扱えなかった複数の次数の遷移確率が混在するような、複雑な分布から生成されるデータも扱えることを示す。

キーワード：隠れマルコフモデル，言語モデル，ノンパラメトリックベイズ，可変オーダー

1. はじめに

隠れマルコフモデルは時系列データを扱うための代表的な手法である。自然言語処理においても、文章を離散の時系列データと見なすことで、単語の品詞推定 [1][2][3] や形態素解析 [4]、機械翻訳の単語アライメントの推定 [5] など、様々な問題に適用されている。

隠れマルコフモデルでは、入力された系列データの各シンボルごとに潜在変数を導入し、潜在クラスの間には依存関係を仮定している。ある位置 t の状態 s_t が1つ前の状態 s_{t-1} のみに依存すると仮定した場合を1次隠れマルコフモデルと呼ぶ。多くの場合、隠れマルコフモデルといえば1次隠れマルコフモデルを指し、このモデルが最も広く利用されている。

しかしながら、現実のデータは必ずしも1次モデルであるとは限らない。例えば、品詞推定では高次隠れマルコフモデルを用いると高い精度で品詞が推定できることが報告されている [6]。

言語データ以外でも、音声認識の音響モデルやバイオインフォマティクスにおける遺伝子やタンパク質の構造解

析、人間の運転行動の分析などで高次の隠れマルコフモデルが応用されている。

実データでは高次隠れマルコフモデルの方が適しているにも関わらず、1次の手法が用いられる理由の1つに計算量があげられる。単純に計算した場合、次数 n に対して計算量は $O(K^{n+1})$ となる。そのため、現実的な計算の都合から扱える次数はせいぜい2次程度となる。これと関連して、データスパースネスの問題があげられる。次数が高くなるにつれて、考慮すべき状態遷移の組み合わせも指数的に増大する。既存の評価データでは多数のパラメータを学習するためには十分なサイズと言えず、高次隠れマルコフモデルを適用しても学習がうまくいかない。

我々は、データスパースネスの問題に対し、状態遷移の事前分布に階層ディリクレ過程を置き、データに合わせた状態数を推定すると同時に適切なスムージングを行うことで対処する。また、系列の各位置毎に次数 n を確率変数として導入し、系列が持つ次数自体もデータから推定する手法を提案する。

2. 先行研究

データからHMMの状態数を推定する手法として、無限の状態数を扱う infinite HMM が提案されている [7][8][9]。Beal ら [7] は状態遷移の事前分布に階層ディリクレ過程を導入することを提案している。ここでは、階層ディリクレ過程によって、新しい状態 $K+1$ への遷移確率を与える

¹ デンソーアイティエラボラトリ
DENSO IT LABORATORY, cross tower 28th Floor, 2-15-1
Shibuya Shibuya-ku Tokyo, 150-0002, Japan

² 統計数理研究所
The Institute of Statistical Mathematics

a) kuchiumi@d-itlab.co.jp

b) daichi@ism.ac.jp

ことで、データに合わせて状態数を可変にしている。Gaelら [8] のアプローチも同様に、遷移確率の事前分布に階層ディリクレ過程を用いているが、階層ディリクレ過程の構築に Chinese Restaurant Process ではなく、Stick-Breaking Process を用いており、加えて状態のサンプリングに Gibbs Sampling でなく動的計画法と Slice Sampling[10] を組み合わせた Beam Sampling を用いることで、高速な学習を行っている。これらの手法は、データから状態数を推定することはできるものの、データを生成したパラメータが持つ真の次数については考慮しておらず、1 次の HMM として扱う。そのため、高次の遷移に対しては新たな状態を生成することで対処し、データが持つ本来の状態数よりも多くの状態が推定される。こうした推定結果は、高次 HMM を 1 次 HMM に展開したものであるため、モデルとしては等価であると考えられるが、多すぎる状態数は人間にとって解釈が難しくなる。

高次 HMM を効率的に学習する手法として、Carter ら [6] は、入力系列に対して n-gram をヒューリスティックでバックオフし、低次で近似する手法を提案している。しかし、事前に次数を決める必要があり、データが持つ真の次数を推定することはできない。また、最尤となる低次の遷移確率に一意に決めてしまうため、初期状態に依存する。この他 Carter ら [6] の手法では状態数は有限としているが、仮に無限とした場合、新しい状態の持つ次数は常に 0 次とされてしまうため、無限次元へ拡張することは簡単ではない。

我々はこの問題に対し、状態だけではなくデータの持つ次数についても確率変数として扱うことで対処する。

3. 提案手法

3.1 生成モデル

以下に、提案手法の生成モデルを (1) に示す。図 1 の隠れマルコフモデルと提案法のグラフィカルモデルで示すように、提案法では各時刻 t での次数 n_t が潜在変数として追加されており、それによって局所的に高次の状態遷移を見る。

$$p(\mathbf{s}, \mathbf{y}, \mathbf{n}) = \prod_{t=1}^T p(s_t | s_{0:t-1}, n_t) p(y_t | s_t) p(n_t | s_{0:t-1}, n_{0:t-1}) \quad (1)$$

$\mathbf{s}$ は状態列 $\mathbf{s} = \{s_0, s_1, \dots, s_T\}$, $\mathbf{y}$ は観測系列 $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$, $\mathbf{n}$ は各時刻の状態の次数 $\mathbf{n} = \{n_0, n_1, \dots, n_T\}$ を表す。

3.2 状態のサンプリング

状態数と次数の推定には、[8] らの Beam Sampling を拡張した手法を用いる。infinite HMM では考慮する状態数が無限のため、そのままでは動的計画法を適用できない。そのため、スライスサンプリングを導入して無限の状態を

有限に制限する。具体的には $n_t > 0$ のとき、以下のよう
に補助変数 μ を導入する。

$$p(\mu_t | s_{0:t}, n_t) = \frac{\mathbb{I}\{0 < \mu_t < p(s_t | s_{0:t-1}, n_t)\}}{p(s_t | s_{0:t-1}, n_t)} \quad (2)$$

補助変数 μ を導入した前向き確率 $p(s_{t-n_t+1:t}, y_{1:t}, \mu_t, n_t)$ を (3) に示す。

$$p(s_{t-n_t+1:t}, y_{1:t}, \mu_t, n_t) = \begin{cases} p(y_t | s_t) \sum_{s_{t-n_{t-1}:t-1}} p(\mu_t | s_{0:t}, n_t) p(s_t | s_{t-n_t:t-1}) p(s_{t-n_{t-1}:t-1}, y_{1:t-1}, \mu_{t-1}, n_{t-1}) & n_t = n_{t-1} \\ p(y_t | s_t) \sum_{s_{t-n_t}} p(\mu_t | s_{0:t}, n_t) p(s_t | s_{t-n_t:t-1}) p(s_{t-n_{t-1}:t-1}, y_{1:t-1}, \mu_{t-1}, n_{t-1}) p(s_{t-n_t:t-n_{t-1}-1}) & n_t > n_{t-1} \\ p(y_t | s_t) \sum_{s_{t-n_t}} p(\mu_t | s_{0:t}, n_t) p(s_t | s_{t-n_t:t-1}) \sum_{s_{t-n_{t-1}:t-n_{t-1}-1}} p(s_{t-n_{t-1}:t-1}, y_{1:t-1}, \mu_{t-1}, n_{t-1}) & n_t < n_{t-1} \end{cases} \quad (3)$$

(3) に従い状態を後ろ向きに状態列をサンプリングすることで、状態の遷移確率を推定する。

(3) では、各時刻 t の次数によって 1 時刻前の前向き確率を適切に周辺化したり、事前分布となる状態の n-gram 確率を掛けることが必要となる。各時刻のデータの次数 n_t については、同時前向き確率を計算することで状態列と同時にサンプリングすることも可能であるが、計算の効率化のために、我々は事前に次数のみをサンプリングする。

3.3 次数のサンプリング

データ $\mathbf{y}$ の確率は、

$$p(\mathbf{y}) = \sum_{\mathbf{n}} \sum_{\mathbf{s}} p(\mathbf{s}, \mathbf{y}, \mathbf{n}) \quad (4)$$

で表される。時刻 t の次数 n_t の事後分布は、ベイズ則より (5) と表すことができ、次数が与えられた時の状態遷移確率と次数の事前確率を用いて、各時刻の n_t をサンプリングすることができる。

$$n_t \sim p(n_t | \mathbf{y}, \mathbf{s}, \mathbf{n}_{-t}) \quad (5)$$

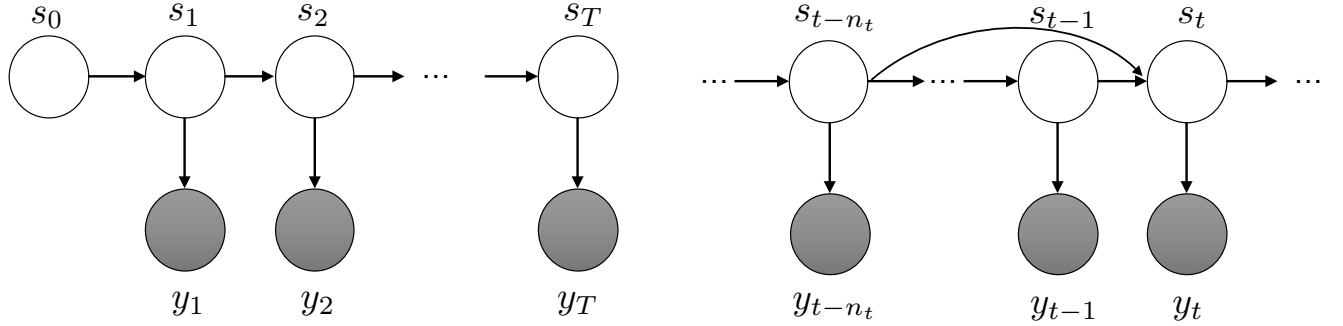
$$\propto p(s_t | \mathbf{n}, \mathbf{y}, \mathbf{s}_{-t}) p(n_t | \mathbf{y}, \mathbf{s}_{-t}) \quad (6)$$

以降で、遷移確率及び次数の事前確率について説明する。

3.4 階層ディリクレ過程

ディリクレ過程では、何らかの基底測度 G_0 を基に離散のディリクレ分布 $G \sim \text{DP}(\alpha, G_0)$ を生成する。ここで、 $\alpha(>0)$ は集中度パラメータを表し、この値が大きいほど G は G_0 に似たものとなる。

隠れマルコフモデルに適用する場合、各状態 s_t が与えられた時の次の状態 s_{t+1} への遷移確率を考える必要がある。各状態遷移確率ごとにディリクレ過程で遷移確率分布を生成した場合、各分布で状態が共有されない。そのため、状態を共有するために各状態遷移ごとに G_0 を共有する。具体的には、基底測度自身もディリクレ過程 $G_0 \sim \text{DP}(\eta, H)$ で生成し、それを共有して次のディリクレ過程 $G_k \sim \text{DP}(\alpha, G_0)$



(a) 1 次 HMM

(b) 提案手法

図 1: 1 次 HMM と提案法のグラフィカルモデル

で各状態線遷移確率を生成する。

DP の構築には、パラメータを陽にする Stick-breaking process (SBP) を用いるか、もしくはパラメータを積分消去して期待値を用いることにより、パラメータを陽にせず観測頻度のみを持つ Chinese restaurant process (CRP) を用いる方法が一般的である。本研究では動的計画法によるサンプリングを用いる都合、一度のサンプリングで複数の状態を新しく生成しやすい SBP を採用した。SBP では、確率分布のパラメータ π を明示的に生成する。具体的には、以下に示すようにベータ分布からサンプルした値を用いて $[0, 1]$ の間で棒を折り、折った棒を G_0 からサンプルした点 k に立て、棒の残りについて再度ベータ分布で折り、次の位置に立てる、という工程を繰り返す。

$$\begin{aligned} \gamma_k &\sim \text{Be}(1, \alpha) \\ \pi_k &= \gamma_k \prod_{j=1}^{k-1} (1 - \gamma_j) \\ G &= \sum_{k=1}^{\infty} \pi_k \delta(\theta_k) \\ \theta_k &\sim G_0 \end{aligned} \quad (7)$$

3.5 階層ディリクレ過程の Chinese District Process 表現

前節で述べた SBP では、確率分布のパラメータを陽に持っている。しかし、本提案手法では状態数 K 及び次数 n についてもデータから決めるため、 K^n 個のパラメータを持つことになる。これは、 K 及び n が大きな値になった場合、現実的ではない。そこで、CRP と同様に、SBP でもパラメータを陽に持たずに観測頻度から期待値を計算する Chinese District Process (CDP) 表現 [9] を用いる。SBP において、ベータ分布からサンプルした値 γ_k で棒を折り、右側を折った棒の残りとしみなした場合、パラメータ π_k は $k-1$ 番目まで右側を選び続け、 k 番目で棒の左を選

んだ時の確率とみなすことができる。棒の左側を選んだ頻度を $n_0(k)$ 、右側を選んだ頻度を $n_1(k)$ とした時、 γ_k の期待値は

$$E[\gamma_k | D] = \frac{1 + n_0(k)}{1 + \alpha + n_0(k) + n_1(k)} \quad (8)$$

と表される。ここで、 D は観測データを表す。CDP では、SBP の棒の折る場所をベータ分布からサンプルする代わりに期待値を用いる。

Pairsley ら [9] は、CDP を隠れマルコフモデルに適用する際に階層化を行わず、状態ごとに独立な遷移確率を生成している。しかし、これではパラメータ間で状態が共有されていない。

SBP を階層化する場合、 $G_k \sim \text{DP}(\alpha, \beta)$ の、 β 自身も DP で生成される。(7) 及び (8) より、 π_k を構成する確率変数 γ の分布は、

$$\gamma_k \sim \text{Be}(\alpha\beta_k, \alpha(1 - \sum_{j=1}^k \beta_j)) \quad (9)$$

であり、 γ_k の期待値は

$$E[\gamma_k | D] = \frac{\alpha\beta_k + n_0(k)}{\alpha\beta_{k:\infty} + n} \quad (10)$$

である。ここで、 n は $n = n_0(k) + n_1(k)$ とした。我々の提案手法では、この階層化した SBP の CDP 表現を用いる。

3.6 次数の事前分布

次数の事前分布には、持橋ら [11] と同様にベータ事前分布を用いる。具体的には、サンプル済みの状態列と次数の観測頻度を基に、以下のようにベータ分布の期待値から計算する。

$$p(n_t = l | \mathbf{y}, \mathbf{s}_{-t}) = \frac{a_l + \epsilon}{a_l + b_l + \epsilon + \zeta} \prod_{i=0}^{l-1} \frac{b_i + \zeta}{a_i + b_i + \epsilon + \zeta} \quad (11)$$

Algorithm 1 学習アルゴリズム

Input: $\mathbf{y} \in \mathbf{Y}$

- 1: Init $\mathbf{S}, \mathbf{N} \sim \text{Uniform}$
- 2: Add $\forall \mathbf{s}, \mathbf{n} \in \mathbf{S}, \mathbf{N}$ to Θ
- 3: **for** $j = 1 \cdots J$ **do**
- 4: **for** s, n in randperm $\mathbf{S}, \mathbf{N}$ **do**
- 5: Remove customers of $\mathbf{s}, \mathbf{n}$ from Θ
- 6: Draw $\mathbf{s}, \mathbf{n}$ according to (5) , (3)
- 7: Add customers of $\mathbf{s}, \mathbf{n}$ to Θ
- 8: **end for**
- 9: Sample hyperparameters of Θ

10: **end for**

3.7 出力確率

出力確率 $p(y_t | s_t)$ については、ディリクレ分布を事前分布とし、以下とした。

$$p(y|s) = \frac{n_{sy} + \phi}{n_s + \phi \cdot V} \quad (12)$$

ここで、 ϕ はディリクレ分布のハイパーパラメータを、 V は語彙数をそれぞれ表す。

3.8 学習アルゴリズム

提案法の学習アルゴリズムを 1 に示す。提案法のサンプリングでは、状態列と次数列を同時にサンプリングしている。そのため、観測されているのは状態 n -gram となる。(10) から分かるように、HDP のパラメータ更新では、サンプリングされたものより低次の状態 n -gram についても更新をする必要がある。そこで、観測された状態 n -gram を基に、確率的に低次の n -gram についても頻度を追加する。ここでは [12] らと同様に、CRP を用いる。我々の HDP 構築には CDP を用いているが、CDP は CRP と等価であるため、観測頻度の更新ではより実装の簡単な CRP を用いた。

3.9 状態と次数の予測

状態と次数の予測は、最尤となる次数と状態列を (13) より予測する。

$$\mathbf{n}^*, \mathbf{s}^* = \underset{\mathbf{n}, \mathbf{s}}{\operatorname{argmax}} p(\mathbf{n}, \mathbf{s}, \mathbf{y}) \quad (13)$$

計算量の問題で単純には計算できないため、(14) に示す通り、予測には動的計画法を用いるのが一般的である。しかし、高次の隠れマルコフモデルの場合は $O(K^n)$ の計算量が必要となる。加えて、提案法の場合は 1 時刻前の次数についても周辺化を行う必要があるため、単純には計算が難しい。そのため、本研究では予測についても学習時と同様に、3.2 で説明した Beam Sampling を用いて、100 epoch のサンプリングを行ない、最初の 10 epoch をバーンインとして除いたサンプルの精度の平均を取ることとした。ただし、パラメータの更新は行わないことに注意されたい。

$$p(n_t, s_{t-n_t}^t, y_0^t) = \sum_{n_{t-1}} \sum_{s_{t-1}} p(y_t | s_t) p(s_t | s_{t-n_t}^{t-1}, n_t) p(n_{t-1}, s_{t-n_{t-1}-1}^{t-1}, y_0^{t-1}) \quad (14)$$

4. 評価

4.1 トイデータによる実験

データの次数及び状態数を推定できることを確認するために、以下に示すパラメータを用意し、そこから 500 件の系列を生成した。

$$\mathbf{A} = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{bmatrix}$$

$$\mathbf{B}_{3-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{B}_{3-2} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{B}_{4-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{B}_{4-2} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{B}_{4-3} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

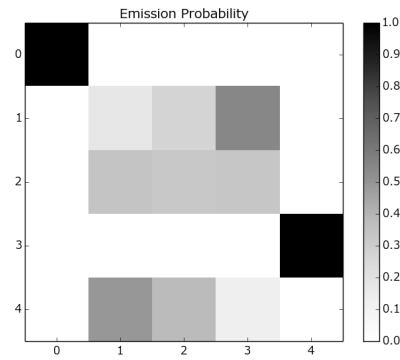
$$\mathbf{E} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$\mathbf{E}$ は出力確率のパラメータ、 $\mathbf{A}$ は 1 次の遷移確率のパラメータ、 $\mathbf{B}$ は 2 次の遷移確率である。 $\mathbf{B}_{3-1}$ は、2 つ前の状態が 3、1 つ前の状態が 1 の時の次の状態への遷移確率を意味する。500 件のデータに対して、提案手法を適用した際の推定結果を図 2a-2g に示す。

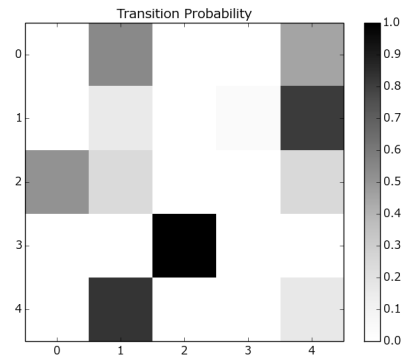
$\mathbf{A}$ を見て分かる通り、状態 4 へは 2 次の遷移が選ばれた時のみ移ることができるように設定している。推定された 1 次の遷移確率 2b を見ると、状態 3 へはどの状態からもほぼ確率がゼロとなっており、2 次の遷移確率 2d-2g から、2-1 または、4-1 という遷移が起こった際に高い確率で状態 3 へ遷移するパラメータが推定されている。出力確率 $\mathbf{E}$ から分かるように、出力シンボル 4 は状態 4 からのみ生成される。推定された出力確率 2a でも、状態 3 からのみ出力シンボル 4 が生成されると推定されており、2 次の状態遷移を通じて到達される真の状態 4 と、推定されたパラメータの状態 3 が対応していることが分かる。

4.2 品詞推定

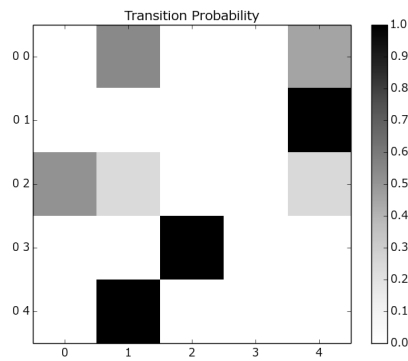
提案手法の評価を行うために、我々は日本語、中国語、英語のそれぞれについて品詞の推定精度を評価する。



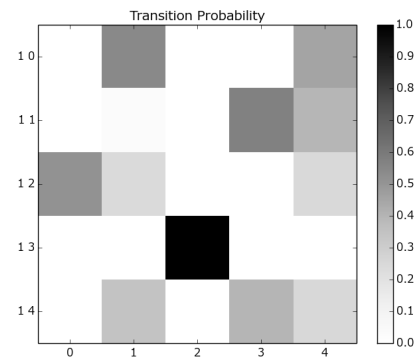
(a) 出力確率



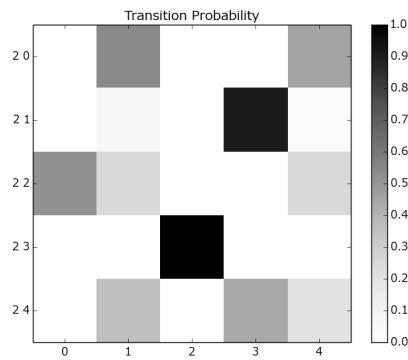
(b) 1 次の遷移確率



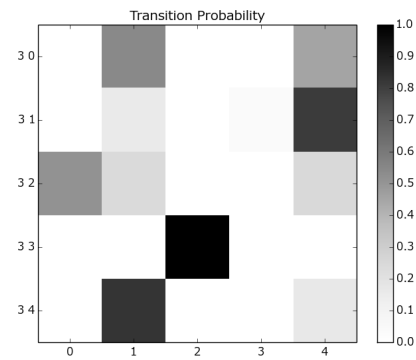
(c) 2 つ前の状態が 0 の時の 2 次の遷移確率



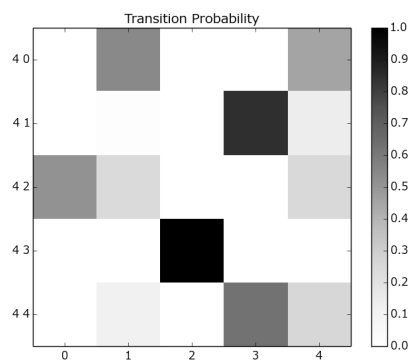
(d) 2 つ前の状態が 1 の時の 2 次の遷移確率



(e) 2 つ前の状態が 2 の時の 2 次の遷移確率



(f) 2 つ前の状態が 3 の時の 2 次の遷移確率



(g) 2 つ前の状態が 4 の時の 2 次の遷移確率

図 2: 推定されたパラメータ

4.2.1 データセット

日本語のデータセットには京大コーパス v4. 0 を、中国語には Chinese Treebank 8.0 を、英語には BNC をそれぞれ用いた。BNC は A, B, C, D, E, F, G, H, J, K の 10 種類のデータからなり、それぞれのデータごとに A0, A1... のようにサブディレクトリに分けられている。ここでは、A0 カテゴリに含まれるデータを用いた。各コーパスの語彙数及び文数を表 1 に示す。

表 1: データセット

コーパス	述べ単語数	サイズ (文)
京大コーパス	1,011,294	37,400
CTB 8.0	1,620,561	71,369
BNC A0	977,097	51,739

我々は各コーパスごとに、無作為に抽出した 10000 文を訓練データとして、1000 文をテストデータとして使用した。

4.2.2 品詞の文脈依存

正解とするコーパスの品詞がどの程度過去の文脈に依存しているかを調べる。各言語のコーパスで、単語が持つ品詞数の分布を図 3 に示す。縦軸は割合、横軸は単語が持つ品詞数のバリエーションを表している。

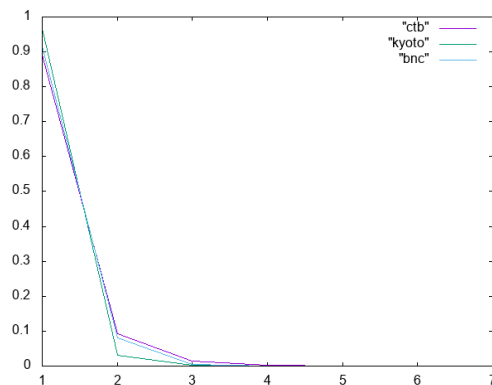


図 3: 単語ごとの取りうる品詞数

図 3 より、どの言語においても多くの単語は 1 つの品詞のみ取りうるということが分かる。つまり、9 割程度の単語については文脈を見る必要がなく、出力確率のみで品詞が推定できることを意味する。

次に、各言語の単語について、品詞の条件付きエントロピー (15) を図 4 に示す。縦軸は条件付きエントロピーを、横軸はいくつ前の品詞まで見るかを意味する。ここで $\mathbf{h}_n$ は n 次の品詞文脈を意味しており、 $n = 0$ の際は文脈を見ない、すなわち品詞 unigram のエントロピーとなる。

$$H(S|\mathbf{h}_n) = - \sum_{\mathbf{h}_n} p(\mathbf{h}_n) \sum_i p(s_i = i|\mathbf{h}_n) \log_2 p(s_i = i|\mathbf{h}_n) \quad (15)$$

図 4 を見ると、どの言語でも品詞の文脈を見ることでエントロピーが大きくなっており、このことからほとんど

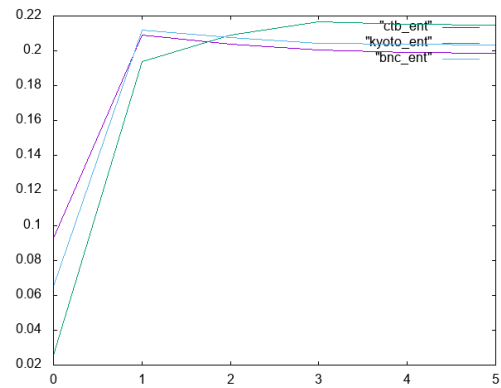


図 4: 品詞の条件付きエントロピー
の単語については文脈を見る必要がないことが伺える。

図 3 及び図 4 から、各言語で 9 割程度の単語については品詞の文脈を見る必要はなく、残りの 1 割の単語についてのみ文脈に依存して品詞が決まることが分かる。では、どのような単語が文脈を考慮すべきなのか。表 2 に、次数に比例して条件付きエントロピーが小さくなる単語の例を示す。どの言語についても品詞の曖昧性のある単語が文脈に依存することが分かる。日本語の場合、接尾辞となりうる単語について、品詞が文脈に依存していることが多い。英語の場合は名詞と動詞や、副詞、形容詞となる単語が文脈に依存している。中国語も同様に、名詞と動詞のどちらにもなりうる単語が文脈に依存している。

表 2: 文脈に依存して品詞が決まる単語の例

言語	単語	取りうる品詞
日本語	より	動詞, 名詞, 副詞, 助詞
	都	名詞, 接尾辞
	だろう	助動詞, 判定詞
	なく	動詞, 接尾辞, 形容詞
	時	名詞, 接尾辞
	中	接頭辞, 接尾辞, 名詞
英語	なり	接尾辞, 助動詞, 助詞, 動詞
	other	SUBST, ADJ
	off	ADJ, PREP, ADV
	much	ADV, ADJ
	need	VERB, SUBST
	play	SUBST, VERB
中国語	no	INTERJ, PREP, ART, VERB, ADV, SUBST
	most	ADJ, ADV
	展	Verb, Noun
	好	Verb, AD, Part, Other
	家	M, Noun
	重要	Verb, Other, Noun
	支持	Noun, Verb
	与	P, Conj
	至	P, Verb, AD, Conj

4.2.3 推定精度

品詞推定の評価は、精度と V-measure[13] を用いた。V-measure はクラスタリングのための評価尺度で、クラスタ内の正解クラスの条件付きエントロピーを表す均一性 (16) と、正解クラスに対するクラスタの条件付きエントロピーを表す完全性 (17) の調和平均で定義される。ここで、 N はデータ数、 n はクラス数、 a_{ij} はクラス j に属するクラス i のデータ数をそれぞれ表す。精度での評価では、正解

クラスに大きな偏りがあった場合、全てのデータを単一のクラスにまとめた時にも高い値となるが、V-measure を用いることで、正解クラスに偏りがある場合でも影響を受けずに評価を行うことができる。

$$h = \begin{cases} 1 & \text{if } H(C, K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & \text{else} \end{cases} \quad (16)$$

$$H(C|K) = - \sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}}$$

$$H(C) = - \sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{ck}}{n} \log \frac{\sum_{k=1}^{|K|} a_{ck}}{n}$$

$$c = \begin{cases} 1 & \text{if } H(K, C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & \text{else} \end{cases} \quad (17)$$

$$H(K|C) = - \sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}}$$

$$H(K) = - \sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{n} \log \frac{\sum_{c=1}^{|C|} a_{ck}}{n}$$

推定された状態と正解コーパスの品詞との対応は、各状態ごとに最頻となる品詞を割り当てることで行なった。

提案法の数値を評価するにあたり、Gael ら [8] の Beam Sampling をベースラインとして用いた。評価結果を表 3 に示す。iHMM がベースラインに該当する。CTB を用いた評価は時間の都合で $n = 2$ までとした。

表 3: 精度と V-measure

dataset	KC	BNC A0	CTB 8.0
iHMM Prec	0.659	0.590	0.544
iHMM V-1	0.383	0.418	0.267
vHMM (n=2) Prec	0.687	0.617	0.592
vHMM (n=2) V-1	0.368	0.473	0.275
vHMM (n=3) Prec	0.671	0.640	-
vHMM (n=3) V-1	0.354	0.469	-

英語と中国語では、考慮する最高次数を大きくすることで、精度の向上が確認できる。一方で、V-measure を比較すると京大コーパスでは 1 次の iHMM の方が最も高い値となっている。iHMM と $n = 2$ の vHMM の均一性 (16) と完全性 (17) を表 4 に示す。比較すると、 $n = 2$ の時は完全性が低くなっている。推定された状態数は、 $n = 1$ の時が 11 であり、 $n = 2, 3$ の時は 20 であった。正解の品詞には京大コーパスの大分類を用いており、品詞数は 14 である。そのため、真の状態数と比べて多く状態が推定された $n = 2, 3$ の場合では完全性が低い値となったと言える。

BNC 及び CTB でも、考慮する次数を大きくすることで精度の向上が確認できた。V-measure では、BNC が $n = 3$ の時に $n = 2$ より低い値となっている。こちらも京大コー

表 4: 京大コーパスの均一性と完全性

	均一性	完全性
n=1	0.428	0.345
n=2	0.461	0.306
n=3	0.440	0.296

パスと同様、推定された状態数が $n = 3$ の時の方が多くなっており、完全性が低くなっていた。図 5 に、日本語の解析結果の例を示す。単語の色は獲得した状態を、「:」の次の数値は推定された次数を意味する。

接尾辞となる「氏」や「中」、「日」は高次を見ることが曖昧性を解消した状態が得られている。特に、「日」は日付を表す場合と「日本」を意味する「日」の場合とが区別されていることが分かる。

4.3 気象の分析

言語以外のデータについても提案手法を適用した際に、どのような結果が得られるかを見るため、我々は気象庁が公開する気象データ*1を用いた。2014 年から 2017 年にかけての 4 年分について、5 月 1 日から 8 月 31 日までの 4 ヶ月の東京の気象データをダウンロードし、このうち 6 時から 18 時の気象概況を月ごとにまとめたものを 1 つの時系列データとして扱った。図 6 に、実験に用いたデータの例を示す。実験で用いる提案手法の最大次数は $n = 4$ とした。図 7 に推定された状態と次数の例を示す。気象概況を色分けすることで状態を、「:」の後ろに付属する数値が次数を表す。

結果を見ると、天気の良い日は主に黄色とマゼンタとなっている。青色で囲われた天候は、前日までの履歴を見て、天候が回復や下降に向かう際の変化途中の状態と考えられ、高次の依存が見られた。

5. 連続値への対応

本報告では、離散のデータのみを対象として評価を行なった。しかし、提案法自体は特にデータを離散に限定していない。(12) では、出力確率にディリクレ分布を仮定したが、連続値を扱う場合にはこの出力確率に連続分布を用いれば良い。データがガウス分布に従う場合は、

$$p(\mathbf{y}|\mathbf{s}) = N(\mathbf{y}; \mu_s, \Sigma_s) \quad (18)$$

$$N(\mathbf{y}; \mu_s, \Sigma_s) = \frac{1}{\sqrt{2\pi}|\Sigma|^{-1}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu_s)^T \Sigma^{-1}(\mathbf{y} - \mu_s)\right) \quad (19)$$

とすればよく、離散値の場合と同様に直接状態列をサンプリングすることでパラメータの更新を行うことができる。

6. まとめ

本稿では、次数を潜在変数として追加した可変次数の無

*1 <http://www.data.jma.go.jp/gmd/risk/obsdl/>

ちょっと:1 前:1 まで:1 は:2 考え:1 られ:1 ない:1 ような:1 変化:1 です:1 。:2
コルジャコフ:1 大統領:1 警護:1 局長:1 ら:1 の:1 名:1 を:1 挙げて:1 いる:3 。:1
首相:1 と:1 山花:1 氏:2 の:2 会談:1 は:1 #:1 日:1 午後:1 、:1 公邸:1 で:1 行わ:1 れる:1 。:1
... 世界:1 が:1 新:1 秩序:1 を:1 求めて:2 いる:1 時期:1 に:1 #:1 世紀:1 を:1 展望:1 する:1 日:2 米:1 の:1 親
密な:1 関係:1 ...
今月:1 中:2 に:1 首相:1 を:1 囲む:2 学者:1 グループ:1 が:1 発表:1 する:1 「:1 村山:1 ビジョン:1 」:1 の:1 基
本:1 的:1 な:1 考え:1 を:1 示した:1 もの:1 。:2

図 5: 日本語の解析結果

薄曇 晴 晴 薄曇 曇一時雨 曇 曇時々雨 曇 ... 曇一時雨
薄曇時々晴 晴 晴 薄曇 快晴 快晴 曇 曇時々雨 ... 曇
曇時々雨一時晴 晴 晴 一時薄曇 晴 一時曇 晴 曇 ... 雨時々曇
曇一時雨 晴 晴 雨時々曇 曇 曇一時雨 晴 薄曇 ... 快晴
晴時々曇 曇 曇 晴一時雨 快晴 曇 晴時々曇 ... 曇
曇時々晴一時雨 曇時々雨 曇一時晴 晴 晴 ... 晴
晴 晴 晴 快晴 晴 曇時々晴 晴 一時曇 ... 曇一時雨
晴 曇時々晴 雨一時曇 薄曇 曇 曇時々雨一時晴 晴 ... 曇
雨時々曇 曇 曇 曇 曇一時晴 曇時々晴 晴 晴 ... 晴時々曇
曇時々雨 曇時々雨 曇 曇 曇 曇 曇時々晴 曇一時雨 ... 雨

図 6: 気象データの例

曇:1 晴:1 快晴:1 薄曇:1 薄曇一時晴:1 曇:1 曇時々晴:3
雨時々曇:1 曇:1 晴:1 雨時々曇:1 曇:1
晴:1 晴:1 晴:1 晴:3 ... 薄曇時々晴:1 曇時々雨:1 曇:1 曇:1
大雨一時曇:1 曇:1 曇時々晴:4
晴:1 晴時々曇:1 雨時々曇:1 曇時々晴:1 雨時々曇:3
曇:1 曇一時雨:1 晴時々曇:1 晴一時曇:1
晴:1 晴時々薄曇:1 晴:1 曇:1 雨:1
雨:1 曇時々晴:3 晴時々曇:1 晴一時薄曇:1 晴時々薄曇:1 薄曇:1

図 7: 気象データの状態と次数

限隠れマルコフモデルの報告を行なった。可変次数無限隠れマルコフモデルでは、必要な箇所のみを高次化することで効率的な計算を行うことができる。加えて従来では扱えなかった次数の異なるパラメータが混在するデータについても扱えることを示した。今後は、言語以外の連続値データについても有効性を調査する予定である。

参考文献

- [1] Goldwater, S. and Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Vol. 45, No. 1, Citeseer, pp. 744–751 (2007).
- [2] Blunsom, P. and Cohn, T.: A Hierarchical Pitman-Yor Process HMM for Unsupervised Part of Speech Induction, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11*, Association for Computational Linguistics, pp. 865–874 (2011).
- [3] Van Gael, J., Vlachos, A. and Ghahramani, Z.: The infi-

- nite HMM for unsupervised PoS tagging, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, Association for Computational Linguistics, pp. 678–687 (2009).
- [4] Uchiumi, K., Tsukahara, H. and Mochihashi, D.: Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, pp. 1774–1782 (2015).
- [5] Vogel, S., Ney, H. and Tillmann, C.: HMM-based word alignment in statistical translation, *Proceedings of the 16th conference on Computational linguistics-Volume 2*, Association for Computational Linguistics, pp. 836–841 (1996).
- [6] Carter, S., Dymetman, M. and Bouchard, G.: Exact sampling and decoding in high-order hidden Markov models, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, pp. 1125–1134 (2012).
- [7] Beal, M. J., Ghahramani, Z. and Rasmussen, C. E.: The infinite hidden Markov model, *Advances in neural information processing systems*, pp. 577–584 (2002).
- [8] Van Gael, J., Saatchi, Y., Teh, Y. W. and Ghahramani, Z.: Beam sampling for the infinite hidden Markov model, *Proceedings of the 25th international conference on Machine learning*, ACM, pp. 1088–1095 (2008).
- [9] Paisley, J. and Carin, L.: Hidden Markov models with stick-breaking priors, *IEEE Transactions on Signal Processing*, Vol. 57, No. 10, pp. 3905–3917 (2009).
- [10] Neal, R. M.: Slice sampling, *Annals of statistics*, pp. 705–741 (2003).
- [11] Mochihashi, D. and Sumita, E.: The infinite Markov model, *Advances in neural information processing systems*, pp. 1017–1024 (2007).
- [12] Teh, Y. W.: A hierarchical Bayesian language model based on Pitman-Yor processes, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 985–992 (2006).
- [13] Rosenberg, A. and Hirschberg, J.: V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure., *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Vol. 7, pp. 410–420 (2007).