一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
IBISML2018-5（2018-06）

可変次数無限隠れマルコフモデル

内海　　慶†　持橋　大地††

† デンソーアイティーラボラトリ　〒 150–0002 東京都渋谷区渋谷 2–15–1
†† 統計数理研究所　〒 190–8562 東京都立川市緑町 10–3

あらまし　従来，時系列データのモデルとして隠れマルコフモデル（HMM）が広く用いられてきた．言語処理では形態素解析や品詞推定，知能化自動車分野では運転行動の分析に用いられており，その用途は多岐に渡る．品詞には高次の依存があることが知られているが，これまでは計算量やデータスパースネスの問題からほとんどの場合，1 次からせいぜい 2 次の HMM しか用いることができなかった．本稿ではこの問題に対処するため，各時刻における次数を潜在変数として追加した無限隠れマルコフモデルの提案を行う．提案手法を用いることで，日本語，英語，中国語の品詞推定で最高精度を達成した．また，運転データの分析を行い，提案手法が連続値データにも有効であることを示した．

キーワード　隠れマルコフモデル，ノンパラメトリックベイズ法

Variable Order Infinite Hidden Markov Models

Kei UCHIUMI† and Daichi MOCHIHASHI††

† DENSO IT LABORATORY,　Shibuya 2–15–1, Shibuya-ku, Tokyo, 150–0002 Japan
†† The Institute of Statistical Mathematics,　Midori-cho 10–3, Tachikawa-shi, 190–8562 Japan

Abstract　Hidden Markov Models have been widely used for modeling of sequential data. For example, it is used for morphological analysis, part-of-speech induction and inference of drive behaviors in the natural language processing and intelligent vehicles, respectively. It is known that the PoS has a higher order dependence. However, because of calculation cost and data sparseness, first order or second order HMM is mostly used. In this paper, to attempt to solve these problems, we propose a new novel infinite HMM which has latent variables to treat its order at each time step. We show that the proposal outperforms conventional methods in PoS induction. As well as discrete dataset, we also show our method can be applied to continuous sequential datasets by analyzing driving logs.

Key words　Hidden Markov Models, Non-parametric Bayes

1. はじめに

従来，隠れマルコフモデル（HMM）は時系列データのモデルとして広く用いられてきた．自然言語処理における形態素解析や品詞推定，知能化自動車における運転行動分析など，適用先は多岐に渡る．自然言語処理の基盤技術である構文解析，固有表現抽出などの教師あり学習では，多くの場合素性として単語や品詞情報を用いており [1] [2] [3] [4] [5] [6]，自然言語処理の応用タスクではコーパスに品詞情報が付与されていることを前提とすることも多い．新しい言語資源にアノテーションを行う場合，全てを人手で行うことは難しく，予めある程度ラベルを付与しておき，専門家によって誤った箇所を修正していく方が効率が良い．

言語は文法を持ち，品詞間には依存関係が存在する．そのため，ラベルなしの言語資源に対しても文法的な規則を考慮した品詞を推定できることが望ましい．この目的には HMM が適しており，実際品詞推定によく用いられている [7] [8] [9]．

HMM は，観測系列 $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$ が背後にある潜在変数列 $\mathbf{s} = \{s_1, s_2, \ldots, s_T\}$ から生成されたとする確率モデルである．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 HMM と呼び，同時確率は，

$$p(\mathbf{s}, \mathbf{y}) = \prod_{t=1}^{T} p(y_t|s_t)p(s_t|s_{t-1})$$

で表される．多くの場合，HMM といえば 1 次 HMM 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，品詞推定では高次 HMM 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 [10]．このことは，品詞には高次の依存関係が存在しており，1 つ前の品詞を見るだけでは不十分なことを示唆している．実データでは高次 HMM 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として，以下が挙げられる．

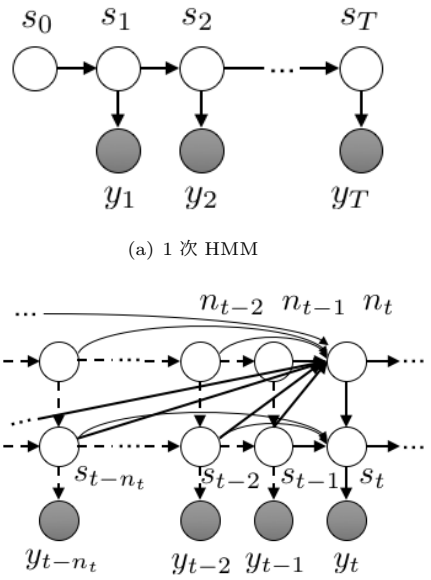
- **真の次数が不明**：対象とする言語の品詞が，いくつ前までの品詞に依存しているのかが分からない．
- **高次モデルの計算量が大きい**：単純に計算した場合，次数 n に対して計算量は $O(K^{n+1})$ となる．そのため，現実的な計算の都合から扱える次数はせいぜい 2 次程度となる．
- **データスパースネスの問題**：次数が高くなるにつれて，考慮すべき状態遷移の組み合わせも指数的に増大する．既存の評価データでは多数のパラメータを学習するためには十分なサイズと言えず，高次 HMM を適用しても学習がうまくいかない．

本論文では，これまで取り組まれて来なかった HMM の次数をデータから推定することに取り組む．加えて，高次化に伴う計算量を抑え，状態遷移に適切な事前分布を置くことでデータスパースネスの問題にも対処する．以降では，各時刻に次数を潜在変数として持つ新たな無限 HMM [11] [12] [9] の提案を行い，品詞推定への適用で従来手法と比較して最高精度を達成したことを報告する．また，連続値のデータへも適用し，非言語データでも有効であることを示す．

2. 先 行 研 究

HMM を用いた品詞推定では，従来モデルを立てる際に予め次数と状態数を決める必要があった [7] [13]．状態については，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 [11] [12] [14]. Beal ら [11] 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ことで，データに合わせて状態数を可変にしている．Gael ら [12] 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 [15] 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 [10] 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，やはり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 [10] 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．



(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

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 $\mathbf{s} = \{s_0, s_1, \ldots, s_T\}$，$\mathbf{y}$ は観測系列 $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$，$\mathbf{n}$ は各時刻の状態の次数 $\mathbf{n} = \{n_0, n_1, \ldots, n_T\}$ を表す．本稿では，式の簡略化のため，時刻 i の状態 s_i から時刻 j の状態 s_j までの状態列をまとめて，$s_i^j = \{s_i, \ldots, s_j\}$ と表記する．

3.2 状態のサンプリング

状態数と次数の推定には，[12] 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 $\mu \sim \mathrm{Beta}(a,b)$ を導入する．[12] では補助変数 μ の確率密度関数に一様分布を仮定しているが，本手法ではベータ分布を用いることで、より効率的に探索を行う．ベータ分布のパラメータを，$a=1, b=1$ とした場合，μ の確率密度関数は一様分布となり，[12] と一致する．

$$p(\mu_t|s_0^t, n_t) = \frac{\mathrm{I}\{0 < \mu_t \leq p(s_t|s_{t-n_t}^{t-1})\}}{p(s_t|s_{t-n_t}^{t-1})} p_\beta(\mu_t; a, b) \quad (2)$$

ここで，$p_\beta(\mu_t; a, b)$ は，ベータ分布の確率密度関数である．補助変数 μ を導入した前向き確率 $p(s_{t-n_t+1}^t|y_1^t, \mu_1^t)$ を (3) に示す．

$$p(s_{t-n_t+1}^t|y_1^t, \mu_1^t) \propto p(s_{t-n_t+1}^t, y_1^t, \mu_1^t) \quad (3)$$

$$= \begin{cases} q \times \sum\limits_{s_{t-n_{t-1}}^{t-n_t}: p_t \geqq \mu_t} p(s_{t-n_{t-1}}^{t-1}, y_1^{t-1}, \mu_1^{t-1}) \text{ if } n_t = n_{t-1} \\ q \times \sum\limits_{s_{t-n_t}: p_t \geqq \mu_t} p(s_{t-n_{t-1}}^{t-1}, y_1^{t-1}, \mu_1^{t-1}) p' \text{ if } n_t > n_{t-1} \\ q \times \sum\limits_{s_{t-n_t}: p_t \geqq \mu_t} \sum\limits_{s_{t-n_{t-1}}^{t-n_t-1}} p(s_{t-n_{t-1}}^{t-1}, y_1^{t-1}, \mu_1^{t-1}) \text{ if } n_t < n_{t-1} \end{cases}$$

読みやすさのため，(3) では $q = p(y_t|s_t)p_\beta(\mu_t; a, b)$，$p_t = p(s_t|s_{t-n_t}^{t-1})$，$p' = p(s_{t-n_t}^{t-n_{t-1}-1})$ とした．(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 $G \sim \mathrm{DP}(\alpha, G_0)$ 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 $G_0 \sim \mathrm{DP}(\eta, H)$ で生成し，それを共有して次のディリクレ過程 $G_k \sim \mathrm{DP}(\alpha, G_0)$ 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．

$$\gamma_k \sim \mathrm{Be}(1, \alpha), \quad \pi_k = \gamma_k \prod_{j=1}^{k-1} (1 - \gamma_j) \quad (7)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k), \quad \theta_k \sim G_0 \ . \quad (8)$$

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 [14] 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

$$\mathrm{E}[\gamma_k|D] = \frac{1 + n_0(k)}{1 + \alpha + n_0(k) + n_1(k)} \quad (9)$$

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 [14] 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，$G_k \sim \mathrm{DP}(\alpha, \beta)$ の，β 自身も DP で生成される．(7) 及び (9) より，π_k を構成する確率変数 γ の分布は，

$$\gamma_k \sim \mathrm{Be}\left(\alpha\beta_k, \alpha\left(1 - \sum_{j=1}^{k} \beta_j\right)\right) \quad (10)$$

であり，γ_k の期待値は

$$\mathrm{E}[\gamma_k|D] = \frac{\alpha\beta_k + n_0(k)}{\alpha\beta_{k:\infty} + n} \quad (11)$$

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．

Algorithm 1 学習アルゴリズム

Require: $\mathbf{y} \in \mathbf{Y}$
Init $\mathbf{S}, \mathbf{N} \sim$ Uniform
Add $\forall \mathbf{s}, \mathbf{n} \in \mathbf{S}, \mathbf{N}$ to Θ
for $j = 1 \cdots J$ **do**
  for s, n in randperm $\mathbf{S}, \mathbf{N}$ **do**
    Remove customers of $\mathbf{s}, \mathbf{n}$ from Θ
    Draw $\mathbf{s}, \mathbf{n}$ according to (5), (3)
    Add customers of $\mathbf{s}$, $\mathbf{n}$ to Θ
  end for
  Sample hyperparameters of Θ
end for

3.6 次数の事前分布

次数の事前分布は，持橋ら [16] と同様に計算する．具体的には，階層ディリクレ過程の各階層に，状態がその階層から生成されるか，より高次の階層から生成されるかを表すベータ分布を導入する．導入したベータ分布を用いて，サンプル済みの状態列と次数の観測頻度を基に，以下のように次数の確率をベータ分布の期待値から計算する．

$$p(n_t = l|\mathbf{y}, \mathbf{s_{-t}}) = \frac{a_l + \nu}{a_l + b_l + \nu + o} \prod_{i=0}^{l-1} \frac{b_i + o}{a_i + b_i + \nu + o} \quad (12)$$

ここで，a_l は観測された l 次の状態数を，b_l は l より高次の状態数を表す．ν 及び o はベータ分布のハイパーパラメータを表す．節 4. で詳細を述べるが，言語では単語が持つ品詞の曖昧性は文脈の次数に対して指数的に減少する．ベータ分布を用いることは，この性質を表現するのに都合が良い．

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．

$$p(y|s) = \frac{n_{sy} + \phi}{n_s + \phi \cdot V} \quad (13)$$

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．(11) からも分かるように，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 [17] 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 (14) より予測する．

$$\mathbf{n}^*, \mathbf{s}^* = \underset{\mathbf{n},\mathbf{s}}{\operatorname{argmax}}\, p(\mathbf{n}, \mathbf{s}, \mathbf{y}) \quad (14)$$

計算量の問題で単純には計算できないため，(15) 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では動的計画法を用いる場合でも $O(K^n)$ 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ず，訓練データから推定したものを用いる．

$$p(s_{t-n_t}^t, y_0^t) = \sum_{n_{t-1}} \underbrace{\sum_{s_{t-1}} \cdots \sum_{s_{t-n_t}}}_{O(K^n)} p(y_t|s_t) p(s_t|s_{t-n_t}^{t-1}) \; p(s_{t-n_{t-1}-1}^{t-1}, y_0^{t-1}) \quad (15)$$

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．

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{A} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$$

$\mathbf{B_{3-1}} = (0\ 0\ 0\ 0\ 1), \mathbf{B_{3-2}} = (0\ 0\ 0\ 0\ 1), \mathbf{B_{4-1}} = (1\ 0\ 0\ 0\ 0)$
$\mathbf{B_{4-2}} = (1\ 0\ 0\ 0\ 0), \mathbf{B_{4-3}} = (1\ 0\ 0\ 0\ 0)$

$\mathbf{E}$ 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 4.1–4.1 に示す．実験では，Beam Sampling に用いる補助変数 μ の確率密度関数を $p_\beta(1/\lambda, \lambda)$ とし，$\lambda = 1$ とした．

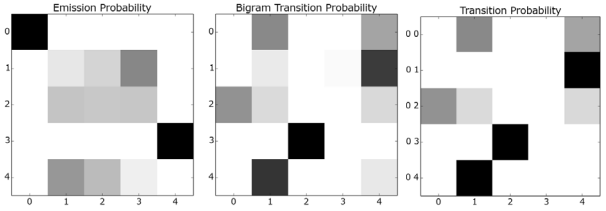
$\mathbf{A}$ を見て分かる通り，状態 4 へは 2 次の遷移が選ばれた時のみ移ることができるように設定している．推定された 1 次の遷移確率 4.1 を見ると，状態 3 へはどの状態からもほぼ確率がゼロとなっており，2 次の遷移確率 4.1–4.1 から，2-1 または，4-1 という遷移が起こった際に高い確率で状態 3 へ遷移するパラメータが推定されている．出力確率 $\mathbf{E}$ から分かるように，出力シンボル 4 は状態 4 からのみ生成される．推定された出力確率 4.1 でも，状態 3 からのみ出力シンボル 4 が生成されると推定されており，2 次の状態遷移を通じ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

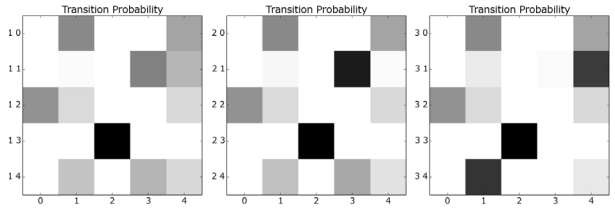
提案手法の評価を行うために，我々は日本語，中国語，英語のそれぞれについて品詞の推定精度を評価する．

4.2.1 データセット

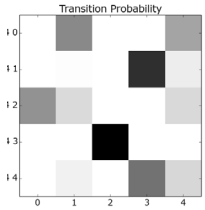
日本語のデータセットには京大コーパス v4．0 を，中国語



(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 1 の時の 2 次の遷移確率

図 2　推定されたパラメータ

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．

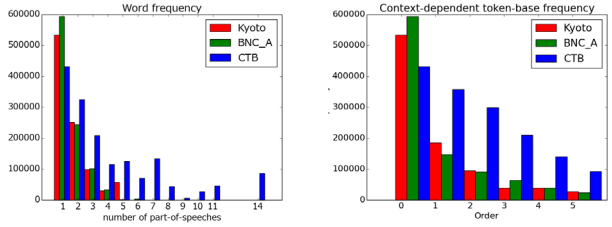
我々は各コーパスごとに，無作為に抽出した 10000 文及び 20000 文を訓練データとして，1000 文をテストデータとして使用した．

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，取り得る品詞数で分けた単語の頻度分布と，文脈の次数に依存して品詞が決まる単語の頻度分布を図 3 に示す．図 3(a) の縦軸は単語の出現頻度，横軸は単語が取り得る品詞の数を表している．図 3(b) の横軸は次数を，縦軸はその次数で品詞が決まる（取りうる品詞数が 1 となる）単語の出現頻度を表す．図 3(a) より，日本語と英語では多くの単語は 1 つの品詞のみ取りうることが分

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



(a) 取り得る品詞数で見た単語の頻度分布　(b) 文脈に依存して品詞が決まる単語の頻度分布

図 3　コーパス中のトークン頻度

かる．中国語でも 1 つのみの品詞を取り得る単語が高頻度となったが，2 つ，3 つの品詞を取る単語の出現頻度も日本語，英語と比較して多くなっている．また，日本語と英語では多くても単語の取り得る品詞は 5 つ程度であるのに対し，中国語では最大 14 の品詞を取り得る単語が存在する．加えて，図 3(b) から分かるように，日本語と英語では次数に応じて単語が持つ品詞の曖昧性が指数的に減少している様子が見れるが，中国語では線形となっている．このことから，中国語の品詞推定は日本語と英語と比較して難しいと言える．

次に，コーパス中の品詞の条件付きエントロピー (16) 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．

$$H(S|\mathbf{h}_n) = -\sum_{\mathbf{h}_n} p(\mathbf{h}_n) \sum_i p(s_t = i|\mathbf{h}_n) \log_2 p(s_t = i|\mathbf{h}_n) \tag{16}$$

図 4 を見ると，どの言語でも高次の文脈を見ることでエントロピーは小さくなっており，高次の品詞遷移が存在することが分かる．

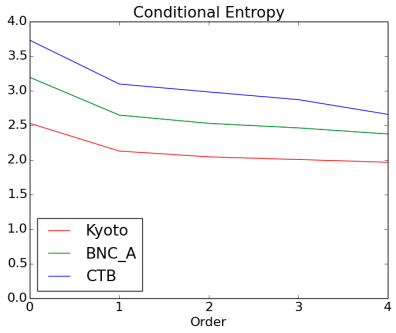


図 4　品詞の条件付きエントロピー

図 3 及び図 4 から，各言語について，多くの単語については品詞の文脈を見る必要はなく，一部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．表 2 に，多数の品詞を取り得る単語の例を示す．こうした単語は文脈に依存して品詞が決まると考えられる．

4.2.3　実験条件

実験では，初期状態数 $K = 10$，出力確率のディリクレ分布

のパラメータ $\phi = 0.1$，遷移確率の階層ディリクレ過程の事前分布のハイパーパラメータ $\alpha = 1$ とした．Beam Sampling の補助変数 μ の確率密度関数のパラメータには，京大コーパスでは $\lambda = 10$，BNC，CTB では $\lambda = 3$ を用いている．λ の値は，各言語ごとに $\lambda = 1, 3, 5, 10$ について予備実験を行い，数値を決めた．提案手法の最大次数は $N = 3$ とした．また，初期状態は初期状態数 K からランダムに初期化した．学習をするにあたり，前処理として各種コーパス中で 2 桁以上の数字については桁数を保持した上で '#' に置き換えている．

いくつかの先行研究では，教師なし品詞推定と言いつつも辞書を用いて品詞の候補を予め制限している [7] [13]．これらは高い品詞推定精度を報告しているが，完全に辞書なし，あるいは事前知識が用いられない場合の評価はしていない．[18] では，辞書などを用いないこれらの手法を比較しており，状態数やハイパーパラメータを調整して大規模なコーパスで学習しても 50%程度の精度であることが報告されている．[8] では上記の辞書などを用いない場合の評価も行なっており，その場合の trigram の DP-HMM で約 68%の精度を達成している．ただし，この場合でも状態数は正しい品詞数と一致するよう事前に与えている．

我々の評価では，上記の事前知識を一切用いていないことに注意されたい．

4.2.4 推 定 精 度

品詞推定の評価は，many-to-1 accuracy(M-1) と V-measure [19] を用いた．M-1 は，予測結果の状態と最も共起する正解品詞クラスをマッピングした後の精度を意味する．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 (17) 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 (18) 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} はクラスタ 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タ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．

$$h = \begin{cases} 1 & \text{if } H(C,K) = 0 \\ 1 - H(C|K)/H(C) & \text{otherwise} \end{cases} \quad (17)$$

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

表 2 品詞曖昧性のある単語の例

言語	単語 (可能な品詞数)
日本語	なり (5), な (5), の (5), あまり (5), 付 (4), 得 (4), なら (4)
英語	like(6),no(6),past(5),down(5), round(5),following(4),opposite(4)
中国語	的 (14), 和 (11), 中 (11), 得 (11), 上 (11), 是 (11), 了 (10)

表 3 M-1 と V-measure. iHMM^1 は 1 次の, iHMM^2 は 2 次の iHMM をそれぞれ表している．

	M-1			V-measure		
Corpus	iHMM^1	iHMM^2	vHMM	iHMM^1	iHMM^2	vHMM
KC 10K	0.661	0.649	**0.718**	0.366	0.320	**0.387**
KC 20K	**0.708**	0.667	0.707	**0.416**	0.376	0.398
BNC 10K	0.591	0.619	**0.707**	0.463	0.481	**0.529**
BNC 20K	0.593	0.631	**0.702**	0.468	0.482	**0.558**
CTB 10K	0.513	0.535	**0.553**	0.242	0.260	**0.287**
CTB 20K	0.490	0.538	**0.575**	0.221	0.278	**0.316**

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

$$c = \begin{cases} 1 & \text{if } H(K,C) = 0 \\ 1 - H(K|C)/H(K) & \text{otherwise} \end{cases} \quad (18)$$

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

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

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の大分類を割り当てることで行った．この際，CTB については Others という大分類に感嘆詞，オノマトペ，名詞修飾，外来語，句読点などがまとめられているため，これらについては細分類のままとした．

提案法の数値を評価するにあたり，Gael ら [12] 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．データセットが異なるため単純には比較できないが, [8] で報告されている DP-HMM は，iHMM で状態数を予め有限とした場合に相当する．

英語，中国語では提案手法がもっとも高い M-1，VM となっている．日本語についても，M-1 については提案法が最も良い結果となった．これは提案法が 3 次の状態遷移を考慮できることからも納得できる．

実験では，2 次の iHMM までをベースラインとして用いている．3 次の iHMM の実験は計算時間の都合で行わなかった．

BNC を用いて，訓練データ 1000 文，テストデータ 1000 文を用いた学習・予測の時間を表 4 に示す．実験に用いた計算機は MacBook Pro 2006（Intel Core i5 3.1Ghz, メモリ 16GB）である．計算速度では，2 次の iHMM よりも 3 次の vHMM の方が早く学習が終わっており，局所的に高次の遷移を扱うことで効率的な学習ができていることが分かる．

表 5 に，1 次の iHMM と 3 次の vHMM の推定結果の例を示す．大まかには 1 次の場合も 3 次の場合も，似た品詞の推定結果となっている．ただし，3 次の場合では「日」,「年」,「人」などの文脈によって名詞と接尾辞となる品詞が上手く推定できている．また，「の」のように文脈によって助詞と名詞の両

表 4 学習・予測に掛かる時間

手法	学習＋予測時間
iHMM (n=2)	5 時間 39 分
vHMM (n=3)	49 分

方になり得る単語についても上手く推定できていた．

表 5　日本語の解析結果の比較

単語	vHMM の品詞	iHMM の品詞	正解品詞
ロシア	名詞	名詞	名詞
軍	名詞	名詞	名詞
は	助詞	助詞	助詞
###	名詞	名詞	名詞
日	接尾辞	名詞	接尾辞
年女	名詞	名詞	名詞
は	助詞	助詞	助詞
推計	名詞	名詞	名詞
で	名詞	助詞	助詞
#####	名詞	名詞	名詞
人	接尾辞	名詞	接尾辞
。	特殊	特殊	特殊
##	名詞	名詞	名詞
年	接尾辞	名詞	接尾辞
から	名詞	助詞	助詞
下がり	動詞	名詞	動詞
新進党	名詞	名詞	名詞
党首	名詞	名詞	名詞
「	名詞	特殊	特殊
海部	名詞	名詞	名詞
俊樹	名詞	名詞	名詞
氏	接尾辞	名詞	接尾辞
」	名詞	名詞	特殊
の	助詞	助詞	助詞
##	名詞	名詞	名詞
%	接尾辞	名詞	接尾辞
を	助詞	助詞	助詞
久保	名詞	名詞	名詞
の	助詞	助詞	助詞
渋面	名詞	名詞	名詞
を	助詞	助詞	助詞
見る	動詞	動詞	動詞
の	名詞	助詞	名詞
は	助詞	特殊	助詞
忍びない	名詞	動詞	形容詞
。	特殊	特殊	特殊

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．(13) 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，

$$p(y|s) = N(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}}\exp\left(-\frac{1}{2}(y-\mu)\Sigma^{-1}(y-\mu)\right)$$

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．

5.1 ガウス分布のパラメータ推定

出力確率にガウス分布を仮定するため，ガウス分布のパラメータについても離散の時と同様にベイズ推定を行う必要がある．サンプルされた状態列から，各状態 k のガウス分布の平均 μ_k の事後分布 (19) を求め，そこから直接 μ_k をサンプリングする．分散についても同様に，精度パラメータ τ についてガンマ事後分布を求め，直接 τ を式 (20) からサンプリングする．ここでは，観測値の D 次元ベクトルの各次元が独立であることを仮定し，分散については全ての状態で共通とした．

$$p(\mu_k|y_1^n, s_1^n, \tau, \mu_0, \rho_0) = N\left(\mu_k|\frac{n_k}{n_k+\rho_0}\hat{y_k} + \frac{\rho_0}{n_k+\rho_0}\mu_0, (\tau(n_k+\rho_0))^{-1}I\right) \tag{19}$$

$$p(\tau|y_1^n, s_1^n, \mu_0, \rho_0, a_0, b_0) = \mathrm{Ga}(\tau|a_n, b_n) \tag{20}$$

$$a_n = a_0 + \frac{n_k D}{2}$$

$$b_n = b_0 + \sum_{k=1}^{K}\left(\frac{1}{2}\sum_{i=1}^{n}\delta(s_i=k)||y_i-\hat{y_k}||^2 + \frac{n_k\rho_0}{2(\rho_0+n_k)}||\hat{y_k}-\mu_0||^2\right)$$

5.2 連続値データへの適用

連続値の系列データでの動作を検証するため，我々は自動車の走行ログを用いて運転行動の状態を推定した．データはロサンゼルスの市街地及び高速道路で計測したもので, データ量及び使用した特徴量は表 6 及び表 7 の通りである． 我々のモデルでは，観測値のガウス分布の平均の事前分布に，平均 0，分散 1 のガウス分布を仮定している．そのため，各特徴量について平均 0，分散 1 となるように規格化した．

提案法では，Beam Sampling を用いて状態を推定する都合上，系列が複数必要になる．そのため，系列を 50 ごとで分割し，113 個の系列として用いた．

初期状態数は 2 とし，最大次数は 5 として実験を行った．表 8 に，獲得した状態と運転シーンを示す．各状態の例を，図 5 に示す．時刻ごとの状態と次数の変化についても，図 6 に示す．走行データの動きは，以下となっている．

（1） 駐車状態
（2） 駐車場を出て左折
（3） 交差点を左折
（4） 高速道路に乗る
（5） 高速道路を降りる
（6） 交差点を右折
（7） 道路を左折して駐車場に入る
（8） 駐車

表 6　走行ログの時間と系列長

時間	系列長	サンプリングレート
9 分 35 秒	5728	100[ms]

表 7　走行ログの特徴量

ID	特徴量
1	速度
2	操舵角
3	アクセル量
4	ブレーキ量
5	加速度センサ (X 軸)
6	加速度センサ (Y 軸)

表 8　推定された状態

状態	運転シーン
1	巡行
2	走行
3	カーブ
4	加減速
5,6	停止・徐行

(a) 駐車　(b) バックで発進（方向転換）

(c) 走行　(d) 加速から巡行

(e) 高速道路に進入（右カーブ）　(f) 加速から巡行

(g) 高速を降りる　(h) 一般道との合流で減速

図 5　推定された状態遷移（最大次数 5）

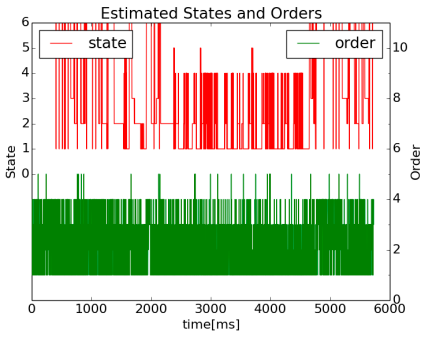


図 6　各時刻の状態と次数

市街地を走っている時と高速道路を走っている時とでは，異なる状態の傾向が得られた．次数については全体を通じてほとんど 1 次から 3 次までの値となっているが，状態の変化が多い時刻では 5 次のような高次の状態遷移が選ばれることが多くなった．

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無限隠れマルコフモデルの報告を行った．提案法では，必要な箇所のみ高次化することで効率的な計算を行うことができ，データに応じた状態数と次数の両方を推定できる．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．

品詞推定での評価では，従来手法を上回る推定精度を達成した．ただし，先行研究と全く同じデータセットを用いた評価は行えていないため，比較のために同じデータセットを用いた評価を行いたい．その他，運転行動データへ適用することで，提案手法が言語のみならず連続値のデータについても適用可能であることを示した．

連続値については数値的な評価は行っていない．今後，適切なタスクにおいて他の手法と比較をする必要がある．

文　献

[1] R. Florian, A. Ittycheriah, H. Jing and T. Zhang: “Named entity recognition through classifier combination”, CoNLL-2003, pp. 168–171 (2003).
[2] H. L. Chieu and H. T. Ng: “Named entity recognition with a maximum entropy approach”, CoNLL-2003, pp. 160–163 (2003).
[3] J. Hammerton: “Named entity recognition with long short-term memory”, CoNLL-2003, pp. 172–175 (2003).
[4] T. Kudo and Y. Matsumoto: “Japanese dependency analysis using cascaded chunking”, CoNLL-2002, pp. 1–7 (2002).
[5] T. Nakagawa: “Multilingual dependency parsing using gibbs sampling”, CoNLL-2007 (2007).
[6] J. Hall, J. Nilsson and J. Nivre: “Single malt or blended? a study in multilingual parser optimization”, Trends in Parsing Technology, pp. 19–33 (2010).
[7] S. Goldwater and T. Griffiths: “A fully bayesian approach to unsupervised part-of-speech tagging”, ACL-2017, pp. 744–751 (2007).
[8] P. Blunsom and T. Cohn: “A hierarchical pitman-yor process hmm for unsupervised part of speech induction”, ACL-HLT-2011, Association for Computational Linguistics, pp. 865–874 (2011).
[9] J. Van Gael, A. Vlachos and Z. Ghahramani: “The infinite hmm for unsupervised pos tagging”, EMNLP-2009, pp. 678–687 (2009).
[10] S. Carter, M. Dymetman and G. Bouchard: “Exact sampling and decoding in high-order hidden markov models”, EMNLP-CoNLL-2012, pp. 1125–1134 (2012).
[11] M. J. Beal, Z. Ghahramani and C. E. Rasmussen: “The infinite hidden markov model”, NIPS-2002, pp. 577–584 (2002).
[12] J. Van Gael, Y. Saatci, Y. W. Teh and Z. Ghahramani: “Beam sampling for the infinite hidden markov model”, ICML-2008, pp. 1088–1095 (2008).
[13] Y. Goldberg, M. Adler and M. Elhadad: “Em can find pretty good hmm pos-taggers (when given a good start)”, ACL-HLT-2008, pp. 746–754 (2008).
[14] J. Paisley and L. Carin: “Hidden markov models with stick-breaking priors”, IEEE Transactions on Signal Processing, pp. 3905–3917 (2009).
[15] R. M. Neal: “Slice sampling”, Annals of statistics, pp. 705–741 (2003).
[16] D. Mochihashi and E. Sumita: “The infinite markov model”, NIPS-2007, pp. 1017–1024 (2007).
[17] Y. W. Teh: “A hierarchical bayesian language model based on pitman-yor processes”, COLING-ACL-2006, pp. 985–992 (2006).
[18] J. Gao and M. Johnson: “A comparison of bayesian estimators for unsupervised hidden markov model pos taggers”, EMNLP-2008, pp. 344–352 (2008).
[19] A. Rosenberg and J. Hirschberg: “V-measure: A conditional entropy-based external cluster evaluation measure.”, EMNLP-CoNLL-2007, pp. 410–420 (2007).