社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－NL－159 (29)
2004／1／14

動的計画法に基づく確率文脈自由文法の変分ベイズ法

栗原 賢一　亀谷 由隆　佐藤 泰介
東京工業大学大学院情報理工学研究科計算工学専攻

確率文脈自由文法の学習法として、I-O アルゴリズム等の EM アルゴリズムが広く知られている。しかし、最尤法である EM アルゴリズムはベイズ学習に比べ訓練データの量が十分でないと過学習するという問題がある。一方、ベイズ学習は困難な積分計算を伴うため従来マルコフ連鎖モンテカルロ法等のサンプリング手法が用いられていたが、計算量が膨大なため大規模な問題への適用が難しかった。近年、マルコフ連鎖モンテカルロ法に比べ効率的な変分ベイズ法が提案され、注目されている。本論文では、確率文脈自由文法に変分ベイズ法を適用し、動的計画法に基づく学習アルゴリズムを導出する。さらに、このアルゴリズムの計算量が I-O アルゴリズムと同じであることを示す。また、変分ベイズ法と EM アルゴリズムの学習精度の比較を行う。

Variational Bayesian Approach to Probabilistic Context-Free Grammar based on Dynamic Programming

Kenichi KURIHARA　Yoshitaka KAMEYA　Taisuke SATO
Department of Computer Science, Tokyo Institute of Technology

The EM algorithm, for example the Inside-Outside algorithm, is a well-known leaning method for probabilistic context-free grammars. However, the EM algorithm is more likely to cause over-fitting than the Bayesian learning when the amount of the training data is not enough. In this paper, we applied the variational Bayesian approach to probabilistic context-free grammars and derived an algorithm based on the dynamic programming approach. In addition, we prove that the computational complexity of our algorithm is equal to that of the Inside-Outside algorithm. We also report results of experiments comparing precisions of the variational Bayesian approach and the EM algorithm.

1 はじめに

確率文脈自由文法 (以下、PCFG) のパラメータの学習アルゴリズムとして Inside-Outside アルゴリズム (以下、I-O アルゴリズム)[2] 等の EM アルゴリズムが広く知られている。しかし、最尤法である I-O アルゴリズムは訓練データの量が十分でないと、過学習する恐れがある。最尤法に対して、ベイズ学習はパラメータを確率変数としパラメータの事後確率分布を学習する。予測分布は事後確率分布を用いた期待値として表わされる。このため、訓練データ量が少ない場合に於いても高い汎化性能が得られる。

従来、ベイズ学習があまり用いられていなかった理由として、事後確率分布の推定に困難な積分計算を伴うということが挙げられる。事後確率分布の計算にはマルコフ連鎖モンテカルロ法等のサンプリング手法が用いられるが、計算量が膨大となり、大規模な問題への適用は困難であった。近年、マルコフ連鎖モンテカルロ法に比べ効率的な変分ベイズ法 [1, 8] が提案され、注目されている。

本論文では、変分ベイズ法を PCFG に適用し事後確率分布の推定式を導出する。さらに、動的計画法に基づく効率的な事後確率分布の推定アルゴリズムを示す。最後に、変分ベイズ法と EM アルゴリズムの学習精度の比較実験を行う。

2 PCFG

2.1 準備

まず、文脈自由言語 (Context Free Grammar 以下、CFG) を 4 項組 $G = (V_N, V_T, R, S)$ により定義する。V_N,V_T,R はそれぞれ、非終端記号、終端記号、生成規則の集合であり、S は開始記号 ($S \in V_N$) で

ある。次に、PCFG を $G(\boldsymbol{\theta})$ により定義する。$\boldsymbol{\theta}$ は $|R|$ 次元のパラメータのベクトルである。生成規則 $r(r \in R)$ のパラメータを $\boldsymbol{\theta}(r)$ により参照する。また、$\boldsymbol{\theta}_A(\alpha)$ という表記により $\boldsymbol{\theta}(A \rightarrow \alpha)$ を表わすことにする。なお、$\sum_\alpha \boldsymbol{\theta}_A(\alpha) = 1$ である。ある非終端記号 A から 0 回以上の生成規則の適用により、単語列 $(w_1, w_2, \ldots, w_n)$ が導出される時、$A \stackrel{*}{\Rightarrow} w_1^n$ と表記する。ある CFG に対して、文 s を導出する全ての適用規則列 $\boldsymbol{r}$ の集合を $\Phi(s)$ と表わすことにする。ある PCFG $G_{\boldsymbol{\theta}}$ が与えられた時、文 s の導出 $\boldsymbol{r}$ による生成確率とは、

$$p(s, \boldsymbol{r}|\boldsymbol{\theta}) = \prod_{r \in R} \boldsymbol{\theta}(r)^{c(r;\boldsymbol{r})} \tag{1}$$

である。ただし、$c(r;\boldsymbol{r})$ とは、生成規則 r が適用規則列 $\boldsymbol{r}$ に出現する回数である。また、ある文 s の生成確率は、

$$p(s|\boldsymbol{\theta}) = \sum_{\boldsymbol{r} \in \Phi(s)} p(s, \boldsymbol{r}|\boldsymbol{\theta}) \tag{2}$$

により、与えられる。

2.2 I-O アルゴリズム

I-O アルゴリズムはチョムスキー標準形 (以下、CNF) である PCFG に対する効率的な EM アルゴリズムである。一般に、EM アルゴリズムはパラメータの更新式を繰り返し適用することによりパラメータを推定する。コーパス C を、$C = (s_1, s_2, \cdots, s_N)$ として時、EM アルゴリズムにより求まる $\boldsymbol{\theta}(A \rightarrow \alpha)$ の更新式は、

$$\hat{\boldsymbol{\theta}}(A \rightarrow \alpha) = \frac{1}{Z_A} \sum_{i=1}^{N} \sum_{\boldsymbol{r} \in \Phi(s_i)} \frac{p(s_i, \boldsymbol{r}|\boldsymbol{\theta})}{p(s_i|\boldsymbol{\theta})} c(A \rightarrow \alpha; \boldsymbol{r}) \tag{3}$$

である。ただし、Z_A は正規化項である。$|\Phi(s_i)|$ が s_i の長さに対して指数関数のオーダであるために、式 (3) の計算量は s_i の長さに対して指数関数のオーダになる。これに対して、動的計画法に基づく I-O アルゴリズムは s_i の長さに対して 3 乗のオーダで式 (3) の計算を行う。

I-O アルゴリズムについて詳しく述べる。まず、$c(r;s)$ を以下により定義する。

$$\begin{aligned} c(r;s) &= \sum_{\boldsymbol{r} \in \Phi(s)} \frac{p(s, \boldsymbol{r}|\boldsymbol{\theta})}{p(s|\boldsymbol{\theta})} c(r;\boldsymbol{r}) \\ &= \frac{1}{p(s|\boldsymbol{\theta})} \sum_{\boldsymbol{r} \in \Phi(s)} c(r;\boldsymbol{r}) \prod_{r \in R} \boldsymbol{\theta}(r)^{c(r;\boldsymbol{r})} \end{aligned} \tag{4}$$

$c(r;s)$ を用いると、式 (3) は以下のようになる。

$$\hat{\boldsymbol{\theta}}(A \rightarrow \alpha) = \frac{1}{Z_A} \sum_{i=1}^{N} c(r; s_i) \tag{5}$$

式 (1) と式 (4) より、次式が得られる。

$$c(r;s) = \frac{\boldsymbol{\theta}(r)}{p(s|\boldsymbol{\theta})} \frac{\partial p(s, \boldsymbol{r}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}(r)} \tag{6}$$

文を $s = (w_1, w_2, \cdots, w_l)$ とし、生成規則 $A \rightarrow BC$ について、$S \stackrel{*}{\Rightarrow} w_1^l$, $A \stackrel{*}{\Rightarrow} w_i^{i+n}$, $B \stackrel{*}{\Rightarrow} w_i^{i+j-1}$, $C \stackrel{*}{\Rightarrow} w_{i+j}^{i+n}$, という導出を考える。この導出の確率は次により与えられる。

$$\begin{aligned} &p(S \stackrel{*}{\Rightarrow} w_1^l, A \stackrel{*}{\Rightarrow} w_i^{i+n}, B \stackrel{*}{\Rightarrow} w_i^{i+j-1}, C \stackrel{*}{\Rightarrow} w_{i+j}^{i+n}) \\ &= p(S \stackrel{*}{\Rightarrow} w_1^{i-1} A w_{i+n+1}^l) \boldsymbol{\theta}(A \rightarrow BC) \\ &\quad \times p(B \stackrel{*}{\Rightarrow} w_i^{i+j-1}) p(C \stackrel{*}{\Rightarrow} w_{i+j}^{i+n}) \end{aligned} \tag{7}$$

したがって、

$$\begin{aligned} \frac{\partial p(s, \boldsymbol{r}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}(A \rightarrow BC)} &= \sum_{n=1}^{l-1} \sum_{i=1}^{l-n} \sum_{j=1}^{n} p(S \stackrel{*}{\Rightarrow} w_1^{i-1} A w_{i+n+1}^l) \\ &\quad \times p(B \stackrel{*}{\Rightarrow} w_i^{i+j-1}) p(C \stackrel{*}{\Rightarrow} w_{i+j}^{i+n}) \end{aligned} \tag{8}$$

である。ここで、外側確率 α、内側確率 β を導入し、

$$\alpha_{i,i+n}(A) = p(S \stackrel{*}{\Rightarrow} w_1^{i-1} A w_{i+n+1}^l) \tag{9}$$

$$\beta_{i,i+j-1}(B) = p(B \stackrel{*}{\Rightarrow} w_i^{i+j-1}) \tag{10}$$

$$\beta_{i+j,i+n}(C) = p(C \stackrel{*}{\Rightarrow} w_{i+j}^{i+n}) \tag{11}$$

とする。これらを用いて、式 (6) は生成規則が $A \rightarrow BC$ $(A, B, C \in V_N)$ の時、

$$\begin{aligned} c(A \rightarrow BC; s) &= \frac{\boldsymbol{\theta}(A \rightarrow BC)}{p(s|\boldsymbol{\theta})} \\ &\quad \times \sum_{n=1}^{l-1} \sum_{i=1}^{l-n} \sum_{j=1}^{n} \alpha_{i,i+n}(A) \beta_{i,i+j-1}(B) \beta_{i+j,i+n}(C) \end{aligned} \tag{12}$$

for $i = 1$ **to** l

$\beta_{i,i}(A) = \boldsymbol{\theta}(A \rightarrow w_i)$

for $n = 1$ **to** $l - 1$

for $i = 1$ **to** $l - n$

$\beta_{i,i+n}(A)$

$$= \sum_{B,C} \boldsymbol{\theta}(A \rightarrow BC) \sum_{j=1}^{n} \beta_{i,i+j-1}(B)\beta_{i+j,i+n}(C)$$

図 1: 内側確率の計算

for $i = 1$ **to** l

for $n = 0$ **to** $l - i$

$\alpha_{i,i+n}(A) = 0$

$\alpha_{1,N}(S) = 1$

for $n = l - 1$ **downto** 1

for $i = 1$ **to** $l - n$

for $k = 1$ **to** n

$\alpha_{i+k,i+n}(C)$

$+= \boldsymbol{\theta}(A \rightarrow BC)\alpha_{i,i+n}(A)\beta_{i,i-1+k}(B)$

$\alpha_{i,i+n-k}(B)$

$+= \boldsymbol{\theta}(A \rightarrow BC)\alpha_{i,i+n}(A)\beta_{i+n+1-k,i+n}(C)$

図 2: 外側確率の計算

となる。また、同様に生成規則が $A \rightarrow a$ ($A \in V_N, a \in V_T$) の時、

$$c(A \rightarrow a; s) = \frac{\boldsymbol{\theta}(A \rightarrow BC)}{p(s|\boldsymbol{\theta})} \sum_{n=1}^{l} \alpha_{i,i}(A) \qquad (13)$$

となる。以上より、内側確率と外側確率は図 1 と図 2 から計算される。

3 変分ベイズ法の PCFG への適用

最尤法がパラメータを点推定するのに対し、ベイズ学習はパラメータを確率変数と見てパラメータの事後確率分布 $p(\boldsymbol{\theta}|D)$(D は訓練データ) を学習する。困難な積分計算を伴うベイズ学習を、変分ベイズ法は近似を用いることにより効率的に計算する [1, 8]。以下では PCFG に変分ベイズ法を適用し、事後確率分布の計算法を導出する。さらに、動的計画法に基づく効率的なアルゴリズムを示す。

3.1 事後確率分布

3.1.1 近似事後確率分布

訓練コーパス C の対数尤度を考え、Jensen の不等式を用いることにより、$\mathcal{F}$ を式 (14) により定義する。ただし、コーパスを C を、$C = (s_1, s_2, \cdots, s_N)$ とし、$\mathcal{R}$ は $\mathcal{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N)$ という形の適用規則列の列であり、$\boldsymbol{r}_i$ は s_i の正解適用規則列である。一般にある文を導出する適用規則列は複数考えられ、正解適用規則列とは、そのうちの真の適用規則列のことである。

$$\begin{aligned}
\mathcal{L}(C) &= \log p(C) \\
&= \log \sum_{\mathcal{R}} \int p(C, \mathcal{R}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \log \sum_{\mathcal{R}} \int q(\mathcal{R}, \boldsymbol{\theta}) \frac{p(C, \mathcal{R}, \boldsymbol{\theta})}{q(\mathcal{R}, \boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\geq \sum_{\mathcal{R}} \int q(\mathcal{R}, \boldsymbol{\theta}) \log \frac{p(C, \mathcal{R}, \boldsymbol{\theta})}{q(\mathcal{R}, \boldsymbol{\theta})} d\boldsymbol{\theta} = \mathcal{F} \qquad (14)
\end{aligned}$$

ここで、$\mathcal{L}(C)$ と $\mathcal{F}$ の差は $p(\mathcal{R}, \boldsymbol{\theta}|C)$ と $q(\mathcal{R}, \boldsymbol{\theta})$ の KL-divergence になっている (式 (15))。

$$\begin{aligned}
\mathcal{L}(C) - \mathcal{F} &= \sum_{\mathcal{R}} \int q(\mathcal{R}, \boldsymbol{\theta}) \log \frac{q(\mathcal{R}, \boldsymbol{\theta})}{p(\mathcal{R}, \boldsymbol{\theta}|C)} d\boldsymbol{\theta} \\
&= KL(q(\mathcal{R}, \boldsymbol{\theta}), p(\mathcal{R}, \boldsymbol{\theta}|C)) \qquad (15)
\end{aligned}$$

C を固定した時 $\mathcal{L}(C)$ は定数であるので、$\mathcal{F}$ を最大化することは $p(\mathcal{R}, \boldsymbol{\theta}|C)$ と $q(\mathcal{R}, \boldsymbol{\theta})$ の KL-divergence を最小化することと等価である。よって、$\mathcal{F}$ を最大化することにより、$p(\mathcal{R}, \boldsymbol{\theta}|C)$ の近似分布 $q(\mathcal{R}, \boldsymbol{\theta})$ を求めることができる。

3.1.2 最適近似事後確率分布の計算

$\mathcal{F}$ を最大化することにより、最適な近似事後確率分布 $q^*(\mathcal{R}, \boldsymbol{\theta})$ の計算法を示す。まず、$q(\mathcal{R}, \boldsymbol{\theta})$ に、$q(\mathcal{R})$ と $q(\boldsymbol{\theta})$ が独立であるという制約を与える (式 (16))。

ただし、$q(\mathcal{R}_i)$ は $\sum_{\boldsymbol{r}\in\Phi(s_i)} q(\mathcal{R}_i=\boldsymbol{r})=1$ を満たす。

$$q(\mathcal{R},\boldsymbol{\theta}) = q(\mathcal{R})q(\boldsymbol{\theta}) = \left\{\prod_{i=1}^{N} q(\mathcal{R}_i)\right\}\left\{\prod_{A\in V_N} q(\boldsymbol{\theta}_A)\right\} \quad (16)$$

また、パラメータ $\boldsymbol{\theta}_A$ の事前分布 $p(\boldsymbol{\theta}_A)$ に共役事前分布である Dirichlet 分布 P_D を仮定し、事前分布 $p(\boldsymbol{\theta})$ を式 (17) により定義する。ただし、$\boldsymbol{u}_A$ は Dirichlet 分布のハイパーパラメータであり、$\Gamma(x)$ はガンマ関数である。

$$p(\boldsymbol{\theta}) = \prod_{A\in V_N} P_D(\boldsymbol{\theta}_A,\boldsymbol{u}_A) \quad (17)$$

$$P_D(\boldsymbol{\theta}_A,\boldsymbol{u}_A) = \frac{1}{Z}\prod_{\alpha;A\to\alpha\in R}\boldsymbol{\theta}_A(\alpha)^{u_{A\to\alpha}-1} \quad (18)$$

$$Z = \frac{\prod_{\alpha;A\to\alpha\in R}\Gamma(u_{A\to\alpha})}{\Gamma\left(\sum_{\alpha;A\to\alpha\in R} u_{A\to\alpha}\right)} \quad (19)$$

$$\boldsymbol{u}_A = \{u_{A\to\alpha}|A\to\alpha\in R\} \quad (20)$$

式 (16) と式 (17) の下、$\mathcal{F}$ に変分法を適用し、$\mathcal{F}$ を最大化する $q(\boldsymbol{\theta}_A)$, $q(\mathcal{R}_i)$ を求める。$q(\mathcal{R}_i)$ は式 (21) のようになる。

$$q(\mathcal{R}_i=\boldsymbol{r}) = \frac{1}{Z_{\mathcal{R}_i}}\prod_{r\in R}\pi(r)^{c(r;\boldsymbol{r})} \quad (21)$$

$$Z_{\mathcal{R}_i} = \sum_{\boldsymbol{r}\in\Phi(s_i)}\prod_{r\in R}\pi(r)^{c(r;t)} \quad (22)$$

$$\pi(A\to\alpha) = \exp\left[\psi(u_{A\to\alpha}) - \psi\left(\sum_{\alpha;A\to\alpha\in R} u_{A\to\alpha}\right)\right] \quad (23)$$

ただし、$\psi(x)$ は digamma 関数であり、以下により定義される。

$$\psi(x) = \frac{\partial\log\Gamma(x)}{\partial x}$$

digamma 関数を含む polygamma 関数の計算機による計算法は文献 [4] に詳しい。

$q(\mathcal{R}_i)$ と同様に、$q(\boldsymbol{\theta}_A)$ も $\mathcal{F}$ に変分法を適用することにより、式 (24) を得る。事前分布 $p(\boldsymbol{\theta}_A)$ 同様、$q(\boldsymbol{\theta}_A)$ も Dirichlet 分布である。

$$q(\boldsymbol{\theta}_A) = \frac{1}{Z_{\boldsymbol{\theta}_A}}\prod_{\alpha;A\to\alpha\in R}\boldsymbol{\theta}_A(\alpha)^{\hat{u}_{A\to\alpha}-1} \quad (24)$$

ただし、$\hat{u}_{A\to\alpha}$ は以下により定義される。

$$\hat{u}_{A\to\alpha} = u_{A\to\alpha} + \sum_{i=1}^{N}\sum_{\boldsymbol{r}\in\Phi(s_i)} q(\mathcal{R}_i=\boldsymbol{r})c(r;\boldsymbol{r}) \quad (25)$$

1. $\boldsymbol{u}^{(0)}$ を用意し、$k\leftarrow 0$ とする。
2. 式 (26) を用いて $\boldsymbol{u}^{(k+1)}$ を計算し、$k\leftarrow k+1$ とする。
3. $\boldsymbol{u}^{(k)}$ が十分収束したら、$\boldsymbol{u}^{*}\leftarrow\boldsymbol{u}^{(k)}$ とする。そうでなければ、2 に戻る。
4. 最適近似事後確率分布 $q^{*}(\boldsymbol{\theta})$ を以下により求める。

$$q^{*}(\boldsymbol{\theta}) = \prod_{A\in V_N} P_D(\boldsymbol{\theta}_A,\boldsymbol{u}_A^{*})$$

図 3: 最適事後確率分布の計算

式 (21) と式 (24) は互いに依存している。これらを交互に適用することにより、最適な近似事後確率分布 $q^{*}(\boldsymbol{\theta})$ を求める。また、各更新において $\mathcal{F}$ は必ず増加するため、収束することが保証されている。$q^{*}(\boldsymbol{\theta})$ を与えるハイパーパラメータ $\boldsymbol{u}^{*}$ は式 (21) と式 (24) より、式 (26) を収束するまで更新することにより求まる。ただし、$\boldsymbol{u}^{(0)}$ は事前分布のハイパーパラメータ $\boldsymbol{u}=\{\boldsymbol{u}_A|A\in V_N\}$ である。

$$u_r^{(k+1)} = u_r + \sum_{i=1}^{N}\sum_{\boldsymbol{r}\in\Phi(s_i)}\frac{c(r;\boldsymbol{r})}{Z_{\mathcal{R}_i}}\prod_{r\in R}\pi^{(k)}(r)^{c(r;\boldsymbol{r})} \quad (26)$$

$$\pi^{(k)}(A\to\alpha) = \exp\left[\psi(u_{A\to\alpha}^{(k)}) - \psi\left(\sum_{\alpha;A\to\alpha\in R} u_{A\to\alpha}^{(k)}\right)\right] \quad (27)$$

図 3 に最適近似事後確率分布 $q^{*}(\boldsymbol{\theta})$ の計算アルゴリズムを示す。

3.2 動的計画法に基づくアルゴリズム

最適近似事後確率分布の計算は式 (26) を用いて、$\boldsymbol{u}^{(k)}$ が繰り返し更新される。しかし、$|\Phi(s_i)|$ が s_i の文長に対して指数関数のオーダであるために、式 (26) の計算量は文長に対して指数関数のオーダである。この困難を克服するために、動的計画法に基づき式 (26) を計算する。

まず、$\gamma(r;s_i)^{(k)}$ を式 (28) により定義する。

$$\gamma(r;s_i)^{(k)} = \sum_{\boldsymbol{r}\in\Phi(s_i)}\frac{c(r;\boldsymbol{r})}{Z_{\mathcal{R}_i}}\prod_{r\in R}\pi^{(k)}(r)^{c(r;\boldsymbol{r})} \quad (28)$$

$\gamma(r;s_i)^{(k)}$ を用いると、式 (26) は次のようになる。

$$u_r^{(k+1)} = u_r + \sum_{i=1}^{N} \gamma(r;s_i)^{(k)} \tag{29}$$

式 (28) は I-O アルゴリズムの式 (4) に対応する。また、π は確率を表わしていないが、式 (4) の $\boldsymbol{\theta}$ に対応する。式 (28) には、2.2 節において示したような式変形を適用することができ、式 (30) になる。

$$\gamma(r;s_i)^{(k)} = \frac{\pi(r)}{Z_{\mathcal{R}_i}} \frac{\partial}{\partial \pi(r)} \sum_{\boldsymbol{r}\in\Phi(s_i)} \prod_{r\in R} \pi(r)^{c(r;\boldsymbol{r})} \tag{30}$$

さらに、外側確率 α、内側確率 β に対応する、μ, ν を用いると $r = A \rightarrow BC$ の時、式 (30) は、

$$\gamma(A\rightarrow BC;s_i)^{(k)} = \frac{\pi(A\rightarrow BC)}{Z_{\mathcal{R}_i}} \times \sum_{n=1}^{l-1}\sum_{i=1}^{l-n}\sum_{j=1}^{n} \mu_{i,i+n}(A)\nu_{i,i+j-1}(B)\nu_{i+j,i+n}(C) \tag{31}$$

となる。同様に、$r = A \rightarrow a$ の時、式 (30) は、

$$\gamma(A\rightarrow a;s_i)^{(k)} = \frac{\pi(A\rightarrow BC)}{Z_{\mathcal{R}_i}} \sum_{n=1}^{l} \mu_{i,i+n}(A) \tag{32}$$

となる。また、$Z_{\mathcal{R}_i}$ は式 (22) より、

$$Z_{\mathcal{R}_i} = \nu_{1,l}(s_i) \tag{33}$$

である。μ, ν の計算は図 2 と図 1 の α を μ に、β を ν に、$\boldsymbol{\theta}$ を π に読み替えることにより可能である。よって、変分ベイズ法の $\boldsymbol{u}$ の更新は I-O アルゴリズムが $\boldsymbol{\theta}$ を更新するのと同じ計算量で計算可能である。

3.3 最尤な適用規則列

最尤法により点推定されたパラメータ $\hat{\boldsymbol{\theta}}$ を用いた最尤な適用規則列は、

$$\hat{\boldsymbol{r}} = \operatorname*{argmax}_{\boldsymbol{r}} p(\boldsymbol{r}|\hat{\boldsymbol{\theta}}) \tag{34}$$

である。

これに対し、変分ベイズ法により求まる最適近似事後確率分布 $q^*(\boldsymbol{\theta}) = \prod_{A\in V_N} P_D(\boldsymbol{\theta}_A, \boldsymbol{u}_A^*)$ を用いた際の最尤な適用規則列は、

$$\begin{aligned}\hat{\boldsymbol{r}} &= \operatorname*{argmax}_{\boldsymbol{r}} \int p(\boldsymbol{r}|\boldsymbol{\theta}) q^*(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \operatorname*{argmax}_{\boldsymbol{r}} \int \prod_{r\in R} \boldsymbol{\theta}(r)^{u_r^*-1+c(r;\boldsymbol{r})} d\boldsymbol{\theta} \\ &= \operatorname*{argmax}_{\boldsymbol{r}} \frac{\prod_{r\in R}\Gamma(u_r^* + c(r;\boldsymbol{r}))}{\prod_{A\in V_N}\Gamma(\sum_{\alpha;A\rightarrow\alpha\in R} u_{A\rightarrow\alpha}^* + c(A\rightarrow\alpha;\boldsymbol{r}))}\end{aligned} \tag{35}$$

である。式 (34) は Viterbi アルゴリズムにより効率的に計算可能であるが、式 (35) には Viterbi アルゴリズムを用いることができない。

表 1: 精度の比較

	0-CB	BT	LT
EM アルゴリズム	0.964	0.762	0.678
変分ベイズ法	0.935	0.814	0.713

4 実験

変分ベイズ法の学習精度と EM アルゴリズム[1]による学習精度を比較した。コーパスには ATR 対話コーパス [9] を用い 5 fold の交差検定を行った。各訓練コーパスには 8,796 文、各テストコーパスには 2,199 文が含まれている. 文法は ATR 対話コーパスに対して開発された文法を CNF に変形したものを用いた。また、ATR 対話コーパスに付加されている正解構文木も CNF に変形した。精度の評価基準には、labeled tree, bracketed tree, zero crossing brackets (以下、LT, BT, 0-CB)[3] を用いた。変分ベイズ法の初期ハイパーパラメータ u_r^0 は任意の生成規則 r に対して $u_r^0 = 2$ とし、EM アルゴリズムの初期パラメータは一様分布とした。変分ベイズ法の終了条件 (図 3 の 3) として、$\sum_{i=1}^{N} \log Z_{\mathcal{R}_i}$ が収束するまでという条件を用いた。

表 1 が実験結果である。0-CB では EM アルゴリズムの方が精度が高いが、より厳しい基準である、BT, LT においては変分ベイズ法の方が精度がよいことが分かる。

なお、3.3 節で述べたように、変分ベイズ法を用いた時の最尤な適用規則列を効率的に計算する方法がない。このため、この実験では全ての可能な適用

[1]具体的には I-O アルゴリズムを用いた。

規則列の集合 $\Phi(s)$ から最尤な適用規則列を探した。$\Phi(s)$ の大きさは文長に対して指数関数のオーダであるが、実験に用いた ATR 対話コーパスでは $\Phi(s)$ はそれほど大きくならなかったため最尤な適用規則列を探すことができた。

5 関連研究

Mackay は変分ベイズ法を隠れマルコフモデルに適用し、隠れマルコフモデルの EM アルゴリズムである Baum-Welch アルゴリズムと同じ計算量のアルゴリズムを導出している。しかし、最尤な状態遷移系列を求める効率的な方法は分からないとしている [6]。

6 まとめ

PCFG に変分ベイズ法を適用し、I-O アルゴリズムと同じ計算量のアルゴリズムを導出した。実験の結果、最尤法である EM アルゴリズムに比較して、変分ベイズ法は BT と LT の基準において精度が向上した。

今後の課題としては、変分ベイズ法において最尤な適用規則列を探す効率的な計算法の研究が挙げられる。

参考文献

[1] Attias, H.: Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. Proc. UAI, 1999.

[2] Baker, J. K.: Trainable grammars for speech recognition. Proceedings of the Spring Conference of the Acoustical Society of America, pp.547-550, 1979.

[3] Goodman, J.: Parsing inside-out, Ph.D. Thesis, Harvard University, 1998.

[4] 石岡 恒憲.: ポリガンマ関数の C 言語, および Fortran77 言語による算譜. 応用統計学, Vol.22, No.1, pp. 23-37, (1993).

[5] 北研二.: 確率的言語モデル. 東京大学出版会, (1990).

[6] Mackay, D.J.C.: Ensemble Learning for Hidden Markov Models. Technical report, University of Cambridge, 1997.

[7] Nigam, K., Mccallum, A., Thrun, S. and Mitchell, T.: The Classification from Labeled and Unlabeled Documents using EM. Machine Learning, 38, pp.103-143, (2000).

[8] 上田修功.: ベイズ学習. 電子情報通信学会誌, Vol. 85, No. 4,6,7,8, 2002.

[9] 浦谷則好, 竹沢寿幸, 松尾秀彦, 森田千帆.: 音声言語データベースの構成 (ATR Integrated Speech and Language Database). ATR Technical Report, TR-IT-0056, エイ・ティ・アール音声翻訳通信研究所, 1994.