自 然 言 語 処 理 126－6
(1998. 7. 23)

スパースな学習データにおける PCFG の確率パラメタの推定法

冨浦 洋一，　日高　達

九州大学大学院システム情報科学研究科

〒812-8581 福岡市東区箱崎 6-10-1

{tom,hitaka}@is.kyushu-u.ac.jp

あらまし　自然言語文の統語構造の曖昧さを絞り込む手法として，統語範疇を意味カテゴリで細分化することにより，係り受け制約を生成規則として表現した確率文脈自由文法を用いる解析が考えられる．しかし，詳細な係り受け制約を記述すると，生成規則数が膨大となり，最尤推定による高信頼度のパラメタ推定値を得るために必要な学習データを収集することが困難となる．本稿では，このような確率文法のパラメタ推定法として，ほとんどの場合に最尤推定量より平均的に誤差が小さく，学習データが十分でない場合により有効となる推定量を提案し，英語の前置詞句の係り先の判定を対象として行なった評価実験について報告する．

キーワード　パラメタ推定，スムージング，確率文脈自由文法，係り受け制約

A Parameter Estimation of a Probabilistic Context Free Grammar on a Sparse Sample

Yoichi TOMIURA, Takeshi NISHIDA and Toru HITAKA

Graduate School of Information Science and Electrical Engineering, Kyushu University

{tom,hitaka}@is.kyushu-u.ac.jp

Abstract　We can disambiguate syntactic structures of a sentence based on a Probabilistic Context Free Grammar (PCFG), where syntactic categories are subdivided semantically so that dependency constraints are expressed in the production rules. But to describe dependency constraints in detail causes an explosion of the number of production rules, which makes it difficult to collect enough size of sample to get a reliable maximum likelihood estimate of parameters in the PCFG. This paper proposes a new estimator of parameters in the PCFG and show the result of an experiment in disambiguation of English prepositional phrase attachment. The mean error of the proposed estimator is practically smaller than the one of the maximum likelihood estimator, and this tendency is more conspicuous on a small size of sample.

key words　Parameter Estimation, Smoothing, PCFG, Dependency Constraints

1 まえがき

自然言語文では，一つの文に対して，与えられた文法によって得られる構文構造は一般に複数存在し，しかもその中には，意味的に不適格なものも多数存在する．可能な構文構造のうち一つを意図して文が作成されたと考えられるため，この構文構造の曖昧さの絞り込みが，構文解析における重要な問題となる．

構文構造の曖昧さの絞り込み手法として，係り受け制約を利用することが考えられる．統語範疇をそれから導出される句の主辞あるいは主辞の意味カテゴリーで細分化して，係り受け制約を確率文脈自由文法の生成規則として表現することができる（確率係り受け文脈自由文法）．この確率文法を用いて解析することにより，可能な構文構造の発生確率を求め，この確率に従って，可能な構文構造の曖昧さの絞り込みを行なうことができる [3][4][5].

しかし，詳細な係り受け制約を記述すると，生成規則数が膨大となり，最尤推定による高信頼度のパラメタ推定値を得るために必要な学習データを収集することが困難となる．本稿では，確率係り受け文脈自由文法のパラメタ推定法として，ほとんどの場合に最尤推定量より平均的に誤差が小さく，学習データが十分でない場合により有効となる推定法を提案し，英語の前置詞句の係り先の判定問題を対象として行なった評価実験について報告する．本手法は，統語範疇を意味カテゴリーで細分化して確率文脈自由文法を精密にした類似研究 [6] に対しても有効と考えられる．

2 確率係り受け文脈自由文法

我々は，係り受け制約を（確率）文脈自由文法の生成規則として表現する手法に関する研究を行なっている [3][4][5]. 本節では，文献 [5] に従って，その概要を説明する．

「昨日買ったリンゴ」における「リンゴ」のように，句の中心的な意味を担う単語をその句の主辞（head）と呼ぶ．句 A が句 B を修飾しているとき，句 A の主辞と句 B の主辞の間に係り受け関係が成立している（句 A の主辞が句 B の主辞に係る）．また，このとき，句 B の主辞の指示対象を句 A の主辞の指示対象で修飾限定するのであるが，限定の仕方，つまり係り受け関係の種類は多様である．一般に句 A の中の語が係り受け関係の種類を規定する情報を持ち，これを句 A の function と呼ぶ*. たとえば,「昨日買ったリンゴを食べる」において，後置詞句「昨日買ったリンゴを」の head は「リンゴ」, function は「を」であり，連体修飾句「昨日買った」の head は「買う」, function は末尾の活用語「た」の活用形「連体」である．また，“jog in a park” の前置詞句 “in a park” の head は “park”, function は “in” である．

単語 α が単語 β に funciton f で規定される関係で係るとき，これが意味的に適格であるためには，α, β, f の間にある一定の制約がある．これを係り受け制約と呼ぶ．文脈自由文法の生成規則で係り受け制約を表現することを考えよう．生成規則,

$$X \longrightarrow Y_1 \cdots Y_i \; X \; Y_{i+1} \cdots Y_l \qquad (1)$$

において，$Y_j (1 \leq j \leq l)$ から導出される句が右辺の X から導出される句を修飾するとする．文脈自由文法において，生成規則 (1) を用いた導出の後，Y_j からの導出と右辺の X からの導出は独立に行なわれるため，Y_j から導出される句の head と右辺の X から導出される句の head の間の係り受け関係が意味的に適格であることを規定できない．そこで，統語範疇を head, function で細分化して，生成規則 (1) を

$$X(h) \longrightarrow Y_1(-h) \cdots X(h) \cdots Y_l(-h) \qquad (2)$$

と

$$Y_j(-h) \longrightarrow Y_j(h_j, f_j) \qquad (j = 1, 2, \cdots, l) \qquad (3)$$

の形式の生成規則とで表現し直す．ここで，$X(h)$ は head が h である統語範疇 X の句を導出する非終端記号であり，$Y(-h)$ は head が h である句に係り得る，統語範疇 Y の句を導出する非終端記号であり，$Y(h, f)$ は head が h, function が f である統語範疇 Y の句を導出する非終端記号である．生成規則 (3) は h_j と h の間に f_j で規定される関係で係る係り受け関係の適格性，つまり，係り受け制約を表している．

【例 1】

$$\begin{aligned} \mathrm{VP}(\text{食べる}) &\longrightarrow \mathrm{PP}(-\text{食べる})\ \mathrm{VP}(\text{食べる}) \\ \mathrm{PP}(-\text{食べる}) &\longrightarrow \mathrm{PP}(\text{リンゴ}, \text{を}) \end{aligned}$$

□

*英語の主格や目的格のように句 A と句 B の位置関係が係り受け関係の種類を規定する場合もある．この場合も位置情報を function とすれば，同様の形式化が可能である．

このように，従来自然言語文法で用いられてきた統語範疇をそれから導出される句の head と function で細分化して係り受け制約を表現した文法を，係り受け文脈自由文法と呼び，これを確率化したものを，確率係り受け文脈自由文法と呼ぶ．

なお，統語範疇をそれから導出される句の head で細分化して確率文法を精密にした類似研究に [6] があるが，この文法においても以下で述べる問題と同様の問題が起こる．

3 確率パラメタの推定法

3.1 最尤推定法を用いた場合の問題点

前節で述べた確率係り受け文脈自由文法の確率パラメタの推定法を考えよう．たとえば，日本語文法で後置詞句（PP）が動詞句（VP）を修飾する場合を考える．九州大学大型計算機センターの公用データベース日本語単語辞書† に登録されている動詞の総数は約 3 万，名詞の総数は約 8 万である．格を規定する格助詞，副助詞は約 10 程度と考えられるので，係り受け制約を表す (3) に対応する以下の形態

$$\mathrm{PP}(-h) \longrightarrow \mathrm{PP}(h', f)$$

の生成規則の総数は膨大になる．頻繁に使われる動詞として 5000 語程度に限定したとしても，一つの動詞に特定の格で係り得る名詞の個数を 50 以上，一つの動詞が取り得る格の種類数を 4 以上と見積もると，少なくとも，$5000 \times 50 \times 4 = 1{,}000{,}000$ 程度の生成規則を考える必要があると思われる．したがって，このままでは，名詞が動詞に係る係り受け制約を表現する規則の適用確率の推定だけを考えても，最尤推定により，十分な信頼度でパラメタ（すなわち規則の適用確率）を推定することができるような大きなサイズの標本（学習データ）を収集することは困難である．

そこで，文献 [5] では，係り受け制約を head そのもののではなく，head の単語の意味カテゴリー（シソーラス上の概念記号）を用いて記述し，上位－下位関係を生成規則として捉えることで，上記の問題を解決している．しかし，このような解決策では，当然，係り受け制約の記述の精度が落ち，また，ある概念の上位概念が必ずしも一つではないことによる問題が起こる．

†見出し語の総数は約 9 万であるが，一つの見出しで複数の品詞を持つものもある．

3.2 提案手法

前節で述べたような問題を解決する手法として，本節では，係り受け制約を記述するのは単語（あるいは，十分に下位の概念）としたままで，最尤推定結果に対するある種のスムージングを行なうことで，係り受け制約を表す生成規則の適用確率の推定を行なう手法を提案する．

確率文脈自由文法 G に基づいて発生したと考えられる構文木列（学習データ）から左辺が B である生成規則のみを重複を許して取り出した標本は，G の生成規則のうち左辺が B である生成規則を確率点（確率変数の定義域）とし，

$$B \longrightarrow \beta_i \quad (i = 1, 2, \cdots, m)$$

の発生確率を G におけるその生成規則の適用確率とする離散分布に従う標本とみなせる．

以下では，一般的に m 個の要素からなる集合 D（便宜上，$D = \{1, 2, \cdots, m\}$ とする）上の離散分布

$$P(t) = \theta_t \qquad (t \in D) \tag{4}$$

$$\begin{cases} \sum_{t \in D} \theta_t = 1, \\ \theta_t \geq 0 \quad (t \in D) \end{cases} \tag{5}$$

を考える．ただし，D 上の任意の 2 点に対して距離が定義されているものとする．

【定義 1】 大きさ N のランダム標本（学習データ）$\boldsymbol{X}$ に対する (4) のパラメータ θ_t の推定量 $\widehat{\theta}_t^S(\boldsymbol{X})$ を以下のように定義する．

$$\widehat{\theta}_t^S(\boldsymbol{X}) = \frac{1}{N} \sum_{i=1}^{N} k(X_i, t)$$

ただし，

$$\begin{cases} \forall x \in D \quad \sum_{t \in D} k(x, t) = 1 \\ \forall x\ y \in D \quad k(x, y) \geq 0 \end{cases} \tag{6}$$

である．□

(6) により，上記定義の推定量は (5) 式を満足する．この手法は，連続分布に対するノンパラメトリックな確率密度の推定法である Parzen 推定法 [1] を，離散分布の確率値の推定のスムージングに利用したものと捉えることができる．

パラメータ θ の推定量 $\widehat{\theta}(\boldsymbol{X})$ が $N \to \infty$ で θ に確率収束するとき，$\widehat{\theta}(\boldsymbol{X})$ を θ の一致推定量と言い，

この性質は，望ましい推定量であるための重要な性質の一つである．

(4) のパラメータ θ_t の最尤推定量 $\widehat{\theta}_t^{ml}(\boldsymbol{X})$ は θ_t の一致推定量である．$\widehat{\theta}_t^{ml}(\boldsymbol{X})$ を定義 1 と同様の形式で表現すると以下のようになる．

【定義 2 】

$$\widehat{\theta}_t^{ml}(\boldsymbol{X}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{X_i,t}$$

ただし，

$$\delta_{x,y} = \begin{cases} 1 & (x = y) \\ 0 & (x \neq y) \end{cases}$$

である． □

したがって，

$$\forall xy \in D \lim_{N\to\infty} k(x,y) = \delta_{x,y} \tag{7}$$

を k に対する制約とすると，$\widehat{\theta}_t^{S}(\boldsymbol{X})$ も θ_t の一致推定量になっていることが分かる．

ここで，

$$\begin{aligned}\widehat{\theta}_t^{S}(\boldsymbol{X}) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{y\in D}\delta_{X_i,y}\cdot k(y,t) \\ &= \sum_{y\in D} k(y,t)\cdot\widehat{\theta}_y^{ml}(\boldsymbol{X})\end{aligned}$$

であることを考慮すると，提案する推定量は，各点 y の発生確率の最尤推定量 $\widehat{\theta}_y^{ml}(\boldsymbol{X})$ の $k(y,t)$ を重みとする平均であることがわかる．後述するように，重み関数 $k(x,y)$ は具体的には，x と y が近いほど大きな値になるように設定する．したがって，t のすぐ近くの点 x の発生確率 θ_x が θ_t と大きく異なるような確率分布に対しては，望ましい結果が期待できない．つまり，提案する推定法により望ましい推定値が得られるには，真の分布関数が局所的に滑らかでなければならない．

3.3 最尤推定量との誤差の比較

前節で提案した推定量は，常に最尤推定量より真値との誤差が小さくなるような推定量ではない．本節では，提案する推定量が最尤推定量より真値との誤差が小さくなるようなパラメタ値がパラメタ空間に占める割合を検討することにより，ほとんどの場合提案する推定量が最尤推定量より平均的に誤差の小さい推定量であり，その傾向は学習データが十分でない場合顕著になることを示す．

$\boldsymbol{X}$ を大きさ N のランダム標本とし，パラメータ θ_t の推定量 $\widehat{\theta}_t(\boldsymbol{X})$ の平均二乗誤差

$$\begin{aligned}MSE(\widehat{\theta}_t(\boldsymbol{X})\,;\,\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}[(\widehat{\theta}_t(\boldsymbol{X})-\theta_t)^2] \\ &= \sum_{\boldsymbol{x}\in D^N}(\widehat{\theta}_t(\boldsymbol{x})-\theta_t)^2 f(\boldsymbol{x}\,;\,\boldsymbol{\theta})\end{aligned}$$

を考え，確率点全体での平均二乗誤差を

$$\sum_{t\in D} MSE(\widehat{\theta}_t(\boldsymbol{X})\,;\,\boldsymbol{\theta})$$

とする．ただし，$f(\boldsymbol{x}\,;\,\boldsymbol{\theta})$ はパラメタ値が $\boldsymbol{\theta}(=\theta_1,\cdots,\theta_m)$ のときの，$\boldsymbol{x}$ の発生確率である．すると，パラメタ値が $\boldsymbol{\theta}$のときに，提案する推定量が最尤推定量より平均的に誤差の小さい推定量であることは，

$$\alpha\cdot\sum_{t\in D} MSE(\widehat{\theta}_t^{ml}(\boldsymbol{X})\,;\,\boldsymbol{\theta}) \geq \sum_{t\in D} MSE(\widehat{\theta}_t^{S}(\boldsymbol{X})\,;\,\boldsymbol{\theta})$$

つまり，

$$\sum_{t\in D}\left[\alpha\cdot MSE(\widehat{\theta}_t^{ml}(\boldsymbol{X})\,;\,\boldsymbol{\theta}) - MSE(\widehat{\theta}_t^{S}(\boldsymbol{X})\,;\,\boldsymbol{\theta})\right] \geq 0 \tag{8}$$

と表せる．これは，正確には，パラメタ値が $\boldsymbol{\theta}$のとき，提案する推定量の確率点全体での平均二乗誤差が最尤推定量の確率点全体での平均二乗誤差の α ($0<\alpha\leq 1$) 倍以下であることを意味する．ただし，$\lim_{N\to\infty}\alpha = 1$ である．パラメタ空間 $\boldsymbol{\Theta}$

$$\boldsymbol{\Theta} = \left\{(\theta_1,\cdots,\theta_m)\,\middle|\,\sum_{t\in D}\theta_t = 1,\quad \theta_t \geq 0(t\in D)\right\}$$

における (8) 式を満たす $\boldsymbol{\theta}$が占める割合が十分 1 に近ければ，ほとんどの場合提案する推定量が最尤推定量より平均的に誤差の小さい推定量であると言える．

簡単のために，定義域 D 上の任意の x，y に対して，

$$d(x,y) = \min(|x-y|, m-|x-y|)$$

が成立するものとし ($d(x,y)$ は x と y の距離)，さらに，以下のように $k(x,y)$ を定義する．

$$k(x,y) = \begin{cases} w & (y = x) \\ \dfrac{1-w}{h} & (y \in D(x\,;\,h)) \\ 0 & (\text{その他}) \end{cases} \tag{9}$$

ただし，$D(x;h)$ は，x の h 番目以内の隣接点から成る集合とする (h は m 未満の偶数，x 自身は含めない)．また，w は $0<w\leq 1$ なる定数で，

$$\lim_{N\to\infty} w = 1 \tag{10}$$

を満たすように設定されるとする．上記の $k(x,y)$ は明らかに，先の条件 (6)(7) を満足する．

このように $k(x,y)$ を定義した場合，各推定量の平均二乗誤差は，以下のように計算できる．

$$
\begin{aligned}
&MSE(\widehat{\theta}_t^{ml}(\boldsymbol{X})\,;\,\boldsymbol{\theta}) \\
&= \sum_{l=0}^{N} \frac{N!}{l!(N-l)!} {\theta_t}^l (1-\theta_t)^{N-l} \left(\frac{l}{N} - \theta_t\right)^2 \\
&= -\frac{1}{N}{\theta_t}^2 + \frac{1}{N}\theta_t \qquad (11)
\end{aligned}
$$

$$
\begin{aligned}
&MSE(\widehat{\theta}_t^{S}(\boldsymbol{X})\,;\,\boldsymbol{\theta}) \\
&= \sum_{\substack{l_1 \geq 0 \\ (l_1+\cdots+l_m=N)}} \cdots \sum_{l_m \geq 0} \left[\frac{N!}{l_1! \cdots l_m!} {\theta_1}^{l_1} \cdots {\theta_m}^{l_m} \left(\frac{1}{N} \sum_{j \in D} k(j,t) l_j - \theta_t \right)^2 \right] \\
&= \frac{1}{N}(\psi_t - {\phi_t}^2) + (\phi - \theta_t)^2 \qquad (12)
\end{aligned}
$$

ただし，

$$
\begin{aligned}
\psi_t &= \sum_{i \in D} k(i,t)^2 \theta_t \\
&= w^2 \theta_t + \left(\frac{1-w}{h}\right)^2 \sum_{i \in D(t\,;\,h)} \theta_i \\
\phi_t &= \sum_{i \in D} k(i,t) \theta_t \\
&= w\theta_t + \frac{1-w}{h} \sum_{i \in D(t\,;\,h)} \theta_i
\end{aligned}
$$

である．

$\boldsymbol{\Theta}$における (8) 式を満たす $\boldsymbol{\theta}$が占める割合を考えるために，$\boldsymbol{\Theta}$における

$$
\alpha \cdot MSE(\widehat{\theta}_t^{ml}(\boldsymbol{X})\,;\,\boldsymbol{\theta}) - MSE(\widehat{\theta}_t^{S}(\boldsymbol{X})\,;\,\boldsymbol{\theta}) \geq 0 \qquad (13)
$$

を満たす $\boldsymbol{\theta}$が占める割合を考える．(11)(12) 式より，

$$
\xi = \theta_t, \qquad \eta = \sum_{i \in D(t\,;\,h)} \theta_i \qquad (14)
$$

とすると，(13) 式は，

$$
a\xi^2 + b\xi\eta + c\eta^2 + d\xi + e\eta \leq 0 \qquad (15)
$$

となる．ただし，

$$
\begin{aligned}
a &= \frac{\alpha - w^2}{N} + (1-w)^2 \\
b &= -\frac{2}{h}\left\{ \frac{w(1-w)}{N} + (1-w)^2 \right\} \\
c &= \left(1 - \frac{1}{N}\right)\left(\frac{1-w}{h}\right)^2 \\
d &= -\frac{\alpha - w^2}{N} \\
e &= \frac{1}{N}\left(\frac{1-w}{h}\right)^2
\end{aligned}
$$

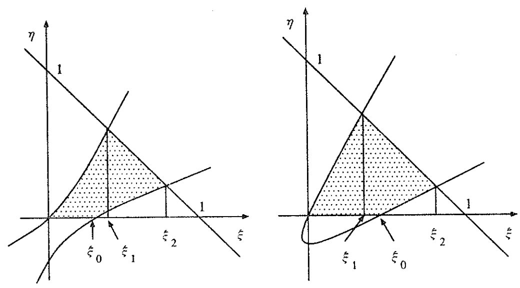
である．$a\xi^2 + b\xi\eta + c\eta^2 + d\xi + e\eta = 0$ は，$N > 1$，$0 < \alpha \leq 1$ で双曲線となる．(15) 式，および，

$$
\xi \geq 0, \quad \eta \geq 0, \quad \xi + \eta \leq 1
$$

を満たす領域（これを $\boldsymbol{\Theta}_{\xi,\eta}$ とする）は，w に対する制限

$$
\alpha > w^2 + \frac{N}{(N-1)h}\left\{ \frac{w(1-w)}{N} + (1-w)^2 \right\} \qquad (16)
$$

の下で，図 1 の (a) あるいは (b) の影の領域となる．



(a) (b)

図 1. 領域 $\boldsymbol{\Theta}_{\xi,\eta}$

したがって，$\boldsymbol{\Theta}$における (13) 式を満たす $\boldsymbol{\theta}$ の占める割合は，

$$
\int_{\boldsymbol{\Theta}_{\xi,\eta}} \frac{(m-1)!\,\eta^{h-1}(1-\xi-\eta)^{m-h-2}}{(h-1)!\,(m-h-2)!} d\xi d\eta
$$

となる．被積分項は $\boldsymbol{\theta}$が $\boldsymbol{\Theta}$上の一様分布の時の (14) 式の (ξ,η) に対する確率密度である．領域の頂点を 4 点とする四角形による $\boldsymbol{\Theta}_{\xi,\eta}$ の近似を用いると，上式は，

$$
(1-\xi_1)^h - (1-\xi_0)^{m-h-1}(1-\xi_2)^h \qquad (17)
$$

と近似できる．ただし，ξ_0 は境界の双曲線が横軸と交わる ξ の値（ただし，原点ではない方）で，ξ_1,ξ_2

は境界の双曲線が $\xi+\eta=1$ と交わる ξ の値（ただし，$\xi_1<\xi_2$）である．(17) 式は，

$$\alpha \leq w \leq \frac{\alpha h+1}{h+1} \tag{18}$$

の範囲内の w で最大となる（$\lim_{N\to\infty}\alpha=1$ であるから，$\lim_{N\to\infty}w=1$ となり，条件 (10) を満たす）．したがって，$\boldsymbol{\Theta}$における (8) 式を満たす $\boldsymbol{\theta}$が占める割合も上記の範囲である程度大きくなることが期待できる．実際，モンテカルロ法による数値計算で，$\boldsymbol{\Theta}$における (8) 式を満たす $\boldsymbol{\theta}$が占める割合（%）を求めてみると，表 1 のようになった．

w \ N	$m/2$	m	$2m$	$5m$	$10m$
0.70	100.00	100.00	99.95	13.34	0.00
⋮	⋮	⋮	⋮	⋮	⋮
0.80	100.00	100.00	99.99	78.10	0.22
⋮	⋮	⋮	⋮	⋮	⋮
0.85	100.00	100.00	100.00	95.34	4.21
0.86	100.00	100.00	100.00	96.95	5.91
0.87	100.00	100.00	100.00	97.69	8.16
0.88	100.00	100.00	100.00	98.24	11.80
0.89	100.00	100.00	100.00	98.68	14.14
0.90	100.00	100.00	100.00	98.76	14.63
0.91	100.00	100.00	100.00	98.56	13.73
0.92	100.00	100.00	100.00	97.74	8.24
0.93	100.00	100.00	100.00	90.71	1.48
0.94	100.00	100.00	99.91	21.58	0.00

表 1. $\boldsymbol{\Theta}$における (8) を満たす $\boldsymbol{\theta}$が占める割合
（α 固定し，w を色々な値にした場合）

ただし，$m=100$，$\alpha=0.9$，$h=8$ である．

$$\frac{\alpha h+1}{h+1} \simeq 0.91$$

であるから，(18) 式の範囲で最大になっていることがわかる．他の m，α，h に対しても同様の傾向が得られた．

また，$m=100$，$h=8$，$w=\alpha$ として，$\boldsymbol{\Theta}$における (8) を満たす $\boldsymbol{\theta}$が占める割合がほぼ一定（約 80%）となるように α の値を調整してみた結果を表 2 に示す．

N	$m/2$	m	$2m$	$5m$	$10m$
α (割合)	0.44 (84.8)	0.59 (83.3)	0.74 (86.7)	0.87 (82.4)	0.93 (82.3)

表 2. $\boldsymbol{\Theta}$における (8) を満たす $\boldsymbol{\theta}$が占める割合
（$w=\alpha$ とし，α を調整した場合）

( ) の中の値は，(8) を満たす $\boldsymbol{\theta}$が占める割合（%）である．これから分かるように，サンプル数 N（厳密には N/m）に応じて，$\lim_{N\to\infty}\alpha=1$ となるように α を適切に選ぶと，ほとんどのパラメタ値において，提案推定量の確率点全体での平均二乗誤差が最尤推定量の確率点全体での平均二乗誤差の α 倍以下となる．しかも，N が小さい場合は，α を小さく設定してもその割合は高いので，サンプル数が小さい場合には，最尤推定と比較して確率点全体での平均二乗誤差が小さいという傾向が顕著になることが分かる．

3.4　係り受け制約を表す規則の適用確率の推定

構文木列から，左辺が $X(-h)$ である生成規則のみを重複を許して取り出した列を $\langle\delta_1,\delta_2,\cdots,\delta_N\rangle$ とする．定義 1 の推定法により，係り受け制約を表す規則

$$X(-h) \longrightarrow X(f,h')$$

の適用確率は，

$$\frac{1}{N}\sum_{i=1}^{N} k(\delta_i,\, X(-h) \longrightarrow X(f,h'))$$

と推定できる．ただし，$k(\delta,\delta')$ は，

$$k(\delta,\delta') = \begin{cases} w & ;\ \delta=\delta' \\ \lambda(\delta)\cdot R^{-d(\delta,\delta')} & ;\ \delta\neq\delta' \end{cases}$$

とする．R は $R>1$ なる適当な定数であり，$d(\delta,\delta')$ は δ とδ' との距離であり，シソーラスを用いて以下のように定義する．

$$d(X(-h)\to X(f_i,h_i),\, X(-h)\to X(f_j,h_j)) = \begin{cases} h_i \text{ と } h_j \text{ のシソーラス上での最短パス長} & ;\ f_i=f_j \\ \infty & ;\ f_i\neq f_j \end{cases}$$

また，$\lambda(\delta)$ は δ に依存する定数で，(6) を満足するように，

$$\lambda(\delta) = \frac{1-w}{\displaystyle\sum_{\delta'\in \mathrm{Rules}(\delta)} R^{-d(\delta,\delta')}}$$

と定める．ここで，$\mathrm{Rules}(\delta)$ は δ と同じ左辺を持つ δ 以外のすべての生成規則からなる集合である．

4　実験

実際の確率係り受け文脈自由文法のパラメタ値に対して，提案推定量の確率点全体での平均二乗誤差が最尤推定量のそれの α $(0<\alpha\leq 1)$ 倍以下にな

る保証はない．しかし，3.3節で見たように，特別な重み関数 $k(x,y)$ に対して，ほとんどのパラメタ領域で，確率点全体での提案推定量の平均二乗誤差が最尤推定のそれの α 倍以下になった．しかも，係り受け制約を表す規則

$$X(-h) \longrightarrow X(f,h')$$

の適用確率は，h' と意味の近い h'' に対する規則

$$X(-h) \longrightarrow X(f,h'')$$

の適用確率に近い考えられ，提案推定量の確率点全体での平均二乗誤差が最尤推定量のそれの α 倍以下になると期待できる．本節では，英語の前置詞句の係り先の判定に確率係り受け文脈自由文法を用いた場合の，提案手法と最尤推定との正解率の比較を行なう．

4.1 実験方法

EDR コーパス [2] から，

他動詞句＋名詞句 1 ＋前置詞＋名詞句 2

なる構造に対して，他動詞句の主辞の動詞 v の単語概念 cv，名詞句 1 の主辞の名詞 n_1 の単語概念 cn_1，前置詞 p，名詞句 2 の主辞の名詞の単語概念 cn_2，および前置詞句の係り先 Cat（v に係る時 V，n_1 に係る時 N）を取り出して

$$\langle cv, cn_1, p, cn_2, Cat\rangle \tag{19}$$

を収集した（総数 11570）．この中に含まれるすべての動詞概念の集合を CV，名詞概念の集合を CN，前置詞の集合を Prep とし，実験に用いた文法の生成規則を，

$$\begin{aligned}
\mathrm{S} &\longrightarrow \mathrm{VIP}(cv),\\
\mathrm{VIP}(cv) &\longrightarrow \mathrm{VIP}(cv)\,\mathrm{PP}(-cv),\\
\mathrm{VIP}(cv) &\longrightarrow \mathrm{VT}(cv)\,\mathrm{NP}(-cv),\\
\mathrm{PP}(-cv) &\longrightarrow \mathrm{PP}(p,cn),\\
\mathrm{PP}(p,cn) &\longrightarrow \mathrm{P}(p)\,\mathrm{NP}(cn),\\
\mathrm{NP}(-cv) &\longrightarrow \mathrm{NP}(cn),\\
\mathrm{NP}(cn) &\longrightarrow \mathrm{NP}(cn)\,\mathrm{PP}(-cn),\\
\mathrm{PP}(-cn_1) &\longrightarrow \mathrm{PP}(p,cn_2),\\
\mathrm{NP}(cn) &\longrightarrow cn,\\
\mathrm{VT}(cv) &\longrightarrow cv,\\
\mathrm{P}(p) &\longrightarrow p
\end{aligned}$$

とする．ただし，各 $cv \in \mathrm{CV}$，各 $cn, cn_1, cn_2 \in \mathrm{CN}$，各 $p \in \mathrm{Prep}$ について，上記のような生成規則があるものとする．

このような生成規則を持つ確率文脈自由文法では，$\langle cv, cn_1, p, cn_2, V\rangle$ の発生確率と $\langle cv, cn_1, p, cn_2, N\rangle$ の発生確率の比は，

$$\frac{Pr(\langle cv, cn_1, p, cn_2, V\rangle)}{Pr(\langle cv, cn_1, p, cn_2, N\rangle)} = \frac{q(cv)\cdot P_{cv}(p, cn_2)}{r(cn_1)\cdot P_{cn_1}(p, cn_2)}$$

となる．ただし，$q(cv)$，$r(cn_1)$，$P_{cv}(p,cn_2)$,$P_{cn_1}(p,cn_2)$ は，それぞれ，生成規則

$$\begin{aligned}
\mathrm{VIP}(cv) &\longrightarrow \mathrm{VIP}(cv)\,\mathrm{PP}(-cv), & (20)\\
\mathrm{NP}(cn_1) &\longrightarrow \mathrm{NP}(cn_1)\,\mathrm{PP}(-cn_1), & (21)\\
\mathrm{PP}(-cv) &\longrightarrow \mathrm{PP}(p,cn_2), & (22)\\
\mathrm{PP}(-cn_1) &\longrightarrow \mathrm{PP}(p,cn_2) & (23)
\end{aligned}$$

の適用確率である．したがって，

$$q(cv)\cdot P_{cv}(p,cn_2) > r(cn_1)\cdot P_{cn_1}(p,cn_2)$$

ならば cv に係り（$Cat = V$），

$$q(cv)\cdot P_{cv}(p,cn_2) < r(cn_1)\cdot P_{cn_1}(p,cn_2)$$

ならば cn_1 に係る（$Cat = N$）と判定できる．

$$q(cv)\cdot P_{cv}(p,cn_2) = r(cn_1)\cdot P_{cn_1}(p,cn_2)$$

の場合は判定不能である．実験では，(19) の形態の 5 つ組から成る 11570 個のデータを 10 等分し，その一つをテストデータ，残りを学習データとして，学習データから適用確率を推定し，テストデータに対して上記の方法で係り先の判定を行ない，テストデータを変えてこれを繰り返し，正解率を求めた（クロスバリデーション）．

(20)(21) の形態の生成規則の適用確率の推定は，任意の動詞概念 cv，cv' に対して $q(cv) = q(cv')$，任意の名詞概念 cn，cn' に対して $r(cn) = r(cn')$ という制約下で，最尤推定により求めた．また，係り受け制約を表す (22) (23) の形態の生成規則の適用確率の推定は，最尤推定と，3.4節で述べた提案手法に基づく推定とで行なった．ただし，$\alpha = w = 0.9$，$R = 2$ とした‡．

‡本来は，$PP(-c)$（c は動詞概念あるいは名詞概念）を左辺とする生成規則の学習データにおける出現頻度で α を変えるべきであるが，簡単のために一定とした．

4.2 実験結果

テストデータ $\langle cv, cn_1, p, cn_2, Cat\rangle$ に対して，$PP(-cv)$ あるいは $PP(-cn_1)$ を左辺とする生成規則が，学習データ中に存在しない場合，提案手法および最尤推定法は，ともに (22) あるいは (23) の生成規則の適用確率が推定できない．このようなテストデータが全体の 46.5 % 存在した．

$PP(-cv)$ および $PP(-cn_1)$ を左辺とする生成規則が，学習データ中に存在する場合に限った結果を表 3 に示す．最尤推定に基づいた場合の正解率が 6.5 % と低いのは，学習データが非常にスパースであるため当然であるが，このような学習データに対しても，提案手法に基づいた場合は 60.6% と比較的高い正解率であった．

	正解率	誤り率	判定不能率
提案手法	60.6	15.9	23.5
最尤推定	6.5	0.7	92.8

表 3. 前置詞句の係り先の判定

前置詞 “of” の場合は，その前置詞句は名詞に係る場合が多く，前置詞 “in” の場合は，その前置詞句は動詞に係る場合が多い．このように，前置詞ごとにどちらに係る場合が多いかを学習データから求め，多い方に係ると一意に判定した場合，正解率 78.3 %，誤り率 21.4 %，判定不能率 0.3 % であった．ただし，比較のために，テストデータは表 3 の場合と同じく，$PP(-cv)$ および $PP(-cn_1)$ を左辺とする生成規則が，学習データ中に存在する場合に限っている．単純なヒューリスティックであるにもかかわらず，提案手法に基づくものよりも高い正解率になっていることがわかる．これは，学習データが余りにも少ないことに原因がある．提案推定量の平均二乗誤差は最尤推定量のそれよりほとんどの場合小さくなると言えるが，学習データが余りに少ない場合は，最尤推定量より誤差は少ないが，精度の高い推定値は期待できない．そこで，$PP(-cv)$ および $PP(-cn_1)$ を左辺とする生成規則の学習データにおける頻度が，10 以上，20 以上，30 以上 の各場合に限って，正解率を求めたものを表 4 に示す．対象となるテストデータの数が少ないので，この結果をそのまま信頼することはできないが，ある程度の数の学習データがあれば，先のヒューリスティックに基づく手法より良い結果が得られると期待できる．

テストデータ	正解率	誤り率	判定不能率
10 以上 (444)	80 (10)	12 (1)	8 (89)
20 以上 (94)	86 (13)	6 (0)	8 (87)
30 以上 (27)	96 (7)	0 (0)	4 (93)

表 4. テストデータを制限した前置詞句の係り先の判定

テストデータの欄の ( ) の中の値はテストデータ数，他の欄の ( ) の中の値は最尤推定の場合の正解率，誤り率，判定不能率．

5 むすび

統語範疇をそれから導出される句の head と function で細分化して，係り受け制約を生成規則として表現した確率係り受け文脈自由文法では，生成規則数が膨大となり，最尤推定による高信頼度のパラメタ推定値を得るために必要な学習データを収集することが困難となる．そこで，ほとんどの場合に最尤推定量より確率点全体での平均二乗誤差が小さい推定量を提案した．さらに，英語の前置詞句の係り先の確率係り受け文脈自由文法に基づいた判定実験により，提案手法の有効性を示した．

参考文献

[1] Keinosuke Fukunaga, Introduction to Statistical Patten Recognition, ACADEMIC PRESS INC., Orlando (1972)

[2] 日本電子化辞書研究所，EDR 電子化辞書仕様説明書，1995

[3] 田辺，冨浦，日高：係り受け制約を含む文脈自由文法，情報処理学会研究報告 Vol.95, No.69 (95-NL-108)，pp.95–101 (1995)

[4] 田辺，冨浦，日高：係り受け制約の文脈自由文法への組み込み法，九州大学大学院システム情報科学研究科報告第 1 巻第 1 号，pp.91–94 (1996)

[5] 田辺，冨浦，日高：係り受け制約を組み込んだ PCFG の評価，九州大学大学院システム情報科学研究科報告第 2 巻第 1 号，pp.93–97 (1997)

[6] W. R. Hogenhout and Y. Matsumoto, “Training Stochastical Grammars on Semantic Categories,” IJCAI’95 Workshop on New Approaches to Learning for Natural Language Processing, pp.65-70 (1995)