スパースな学習データにおける確率係り受け文脈自由文法の確率パラメータの推定法

冨浦洋一[†] 日高達[†]

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，最尤推定によって高信頼度のパラメータ推定値が得られる程度に大きなサイズの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ータ推定法として，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った評価実験について報告する．

A Parameter Estimation of PCFG Expressing Dependency Constraints on a Sparse Sample

YOICHI TOMIURA[†] and TORU HITAKA[†]

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 as 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 shows 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 sized sample.

1. まえがき

自然言語文では，1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1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うことができる[6)~8)]．

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，最尤推定によって高信頼度のパラメータ推定値が得られる程度に大きなサイズの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ータ推定法として，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った評価実験について報告する．本手法は，語の共起制約を確率モデルに取り込んだ類似研究[2),5)]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っている[6)~8)]．本章では，文献8)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，

† 九州大学大学院システム情報科学研究科
Graduate School of Information Science and Electrical Engineering, Kyushu University

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headは“jog”，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われるため，Y_jから導出される句のheadおよびfunction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

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

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である句に係りうる，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．たとえば，「食べる」をheadとする動詞句（VP）に後置詞句（PP）が係ることを表現する生成規則，および，「リンゴ」をhead，「を」をfunctionとする後置詞句（PP）が「食べる」に係りうる後置詞句であることを表現する生成規則は，それぞれ以下のようになる．

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

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．

3. 確率パラメータの推定法

3.1 最尤推定法を用いた場合の問題点

前章で述べた確率係り受け文脈自由文法の確率パラメー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式(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，1つの動詞に特定の格で係りうる名詞の個数を50以上，1つの動詞が取りうる格の種類数を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，名詞が動詞に係る係り受け制約を表す生成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ータ（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8)では，係り受け制約をheadそのもののものではなく，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とらえることで，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1つではないことによる問題が起こる．

☆ 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．

☆☆ 見出し語の総数は約9万であるが，1つの見出しで複数の品詞を持つものもある．

なお，語の共起制約を確率モデルに取り込んだ類似研究[2),5)]においても同様の問題が起こると考えられ，これらのモデルのパラメータ推定に関しても本稿で提案する手法は有効である．

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，生成規則

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

の G における適用確率をその生成規則（確率点）の発生確率とする離散分布に従う標本と見なせる．

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

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

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

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）$\boldsymbol{X}$ $(= \langle X_1, X_2, \cdots, X_N \rangle)$ に対する式 (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，

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

である． □

式 (6) により，上記定義の推定量は式 (5) 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とらえることができる．

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 の一致推定量といい，この性質は，望ましい推定量であるための重要な性質の 1 つである．

式 (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．

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

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分かる．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，つね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誤差が小さくなるようなパラメータ値がパラメー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

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

を考え，確率点全体での平均二乗誤差を

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

とする．ただし，$f(\boldsymbol{x}\,;\,\vec{\theta})$ はパラメータ値が $\vec{\theta}$ $(=(\theta_1,\theta_2,\cdots,\theta_m))$ のときの，$\boldsymbol{x}$ の発生確率である．すると，パラメータ値が $\vec{\theta}$ 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，

$$\alpha\cdot\sum_{t\in D}\mathrm{MSE}(\widehat{\theta}_t^{ml}(\boldsymbol{X})\,;\,\vec{\theta}) \geq \sum_{t\in D}\mathrm{MSE}(\widehat{\theta}_t^{S}(\boldsymbol{X})\,;\,\vec{\theta}) \tag{8}$$

と表せる．これは，正確には，パラメータ値が $\vec{\theta}$ 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 $\alpha\,(0<\alpha<1)$ 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．パラメータ空間 $\vec{\Theta}$

$$\left\{(\theta_1,\cdots,\theta_m)\;\middle|\;\sum_{t\in D}\theta_t=1,\ \forall t\in D\ \ \theta_t\geq 0.\right\}$$

における式 (8) を満たす $\vec{\theta}$ 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いえる．

簡単のために，定義域 D を**図 1** のように，円周上の等間隔に並んだ m 個の点の集合とし，D 上の x，y の距離を $\min(|x-y|,m-|x-y|)$ と定義する．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 未満の自然数，$D(x;h)$ に x 自身は含めない）．また，w は以下を満たすように設定されるものとする．

$$\begin{cases} 0<w\leq 1\,,\\ \lim\limits_{N\to\infty} w=1\,. \end{cases} \tag{10}$$

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．

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

ただし，

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

である．したがって，

$$\eta_t=\sum_{i\in D(t\,;\,h)}\theta_i$$

と表すと，式 (8) は，

$$\sum_{t=1}^{m}(a{\theta_t}^2+b\theta_t\eta_t+c{\eta_t}^2+d\theta_t+e\eta_t)\leq 0 \tag{11}$$

となる．ただし，

$$a=\frac{\alpha-w^2}{N}+(1-w)^2\,, \tag{12}$$

$$b=-\frac{2}{h}\left\{\frac{w(1-w)}{N}+(1-w)^2\right\}\,, \tag{13}$$

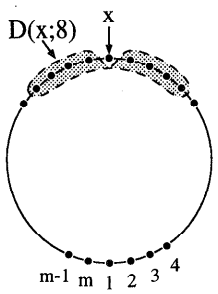


図 1 定義域 D と $D(x\,;\,h)$
Fig. 1 Domain D and $D(x\,;\,h)$.

$$c = \left(1 - \frac{1}{N}\right)\left(\frac{1-w}{h}\right)^2 , \tag{14}$$

$$d = -\frac{\alpha - w^2}{N} , \tag{15}$$

$$e = \frac{1}{N}\left(\frac{1-w}{h}\right)^2 \tag{16}$$

である．ここで，

$$\sum_{t=1}^{m} \theta_t = 1 ,$$

$$\sum_{t=1}^{m} \eta_t = h ,$$

$$\sum_{t=1}^{m} {\theta_t}^2 = \left(\sum_{t=1}^{m} \theta_t\right)^2 - \sum_{t=1}^{m} \sum_{j=1, j\neq t}^{m} \theta_t \theta_j$$

$$= 1 - \sum_{t=1}^{m} \theta_t (\eta_t + \xi_t)$$

$$(\xi_t = 1 - \theta_t - \eta_t)$$

$$= 1 - \sum_{t=1}^{m} \theta_t \eta_t - \sum_{t=1}^{m} \theta_t \xi_t ,$$

$$\sum_{t=1}^{m} {\eta_t}^2 = \left(\sum_{t=1}^{m} \eta_t\right)^2 - \sum_{t=1}^{m} \sum_{j=1, j\neq t}^{m} \eta_t \eta_j$$

$$= h^2 - \sum_{t=1}^{m} \sum_{j=1, j\neq t}^{m} \eta_t \eta_j$$

であるから，式 (11) は，

$$A \sum_{t=1}^{m} \theta_t \xi_t + B \sum_{t=1}^{m} \theta_t \eta_t + C \sum_{t=1}^{m} \sum_{j=1, j\neq t}^{m} \eta_t \eta_j \geq 1 \tag{17}$$

と等価である．ただし，

$$A = \frac{a}{a + ch^2 + d + eh} , \tag{18}$$

$$B = \frac{a - b}{a + ch^2 + d + eh} , \tag{19}$$

$$C = \frac{c}{a + ch^2 + d + eh} \tag{20}$$

である．式 (18)，(19)，(20) に式 (12)，(13)，(14)，(15)，(16) を代入すると，以下が分かる．

- 式 (18) は，$w = \alpha$ で最大（$w \leq \alpha$ で単調増加，$w \geq \alpha$ で単調減少）．
- 式 (19) は，$w = (1 + h\alpha)/(1 + h)$ で最大（$w \leq (1+h\alpha)/(1+h)$ で単調増加，$w \geq (1+h\alpha)/(1+h)$ で単調減少）．
- 式 (20) は w に依存しない．

θ_i，η_i，ξ_i は任意の i で非負であるから，パラメータ空間 $\vec{\Theta}$ における式 (17)（すなわち，$k(x, y)$ を式 (9) のように定めた場合の式 (8)）を満足する $\vec{\theta}$ の割合は，

$$\alpha \leq w \leq \frac{\alpha h + 1}{h + 1} \tag{21}$$

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）．

モンテカルロ法による数値計算で，$\vec{\Theta}$ における式 (17) を満たす $\vec{\theta}$ が占める割合（%）を求めてみると，**表 1** のようになった．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，式 (21) の範囲で最大になっていることが分かる．他の m，h，α 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，$\vec{\Theta}$ における式 (17) を満たす $\vec{\theta}$ 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．( ) の中の値は，式 (17) を満たす $\vec{\theta}$ が占める割合（%）である．これから分かるように，N が小さ

表 1 $\vec{\Theta}$ における式 (17) を満たす $\vec{\theta}$ が占める割合
Table 1 The rate of $\vec{\Theta}$ which the set of $\vec{\theta}$ satisfying Eq. (17) accounts for.

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

($m = 100$, $h = 8$, $\alpha = 0.9$)

表 2 $\vec{\Theta}$ における式 (17) を満たす $\vec{\theta}$ が占める割合を約 80%とする α
Table 2 α when the set of $\vec{\theta}$ satisfying Eq. (17) accounts for about 80% of $\vec{\Theta}$.

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

($m = 100$, $h = 8$, $w = \alpha$)

The value in ( ) is the exact rate of $\vec{\Theta}$ which the set of $\vec{\theta}$ satisfying Eq. (17) accounts for.

い場合は，α を小さく設定しても $\vec{\Theta}$ における式 (17) を満たす（すなわち，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）$\vec{\theta}$ が占める割合は高いので，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．

式 (4) の確率分布関数 P の推定量 $\widehat{P}$ の良さの尺度として，Kullback-Leibler 情報量 $D(P||\widehat{P})$

$$D(P||\widehat{P}) = \sum_{t \in D} P(t) \log \frac{P(t)}{\widehat{P}(t)}$$

がある．これは，2つの確率分布 P と $\widehat{P}$ とのある種の距離を示すもので，この値が小さい方が，$\widehat{P}$ が P に近い，すなわち，良い推定量であるといえる．しかし，ある $i\,(\in D)$ に対して，$P(i) > 0$ かつ $\widehat{P}(i) = 0$ となる場合（標本が十分には大きくないことが原因），$D(P||\widehat{P}) = \infty$ となってしまう．したがって，P の2つの推定量 $\widehat{P_1}$，$\widehat{P_2}$ に対して，$D(P||\widehat{P_1}) = D(P||\widehat{P_2}) = \infty$ であるということで，両推定量の精度が等しいとか，$D(P||\widehat{P_1}) = \infty$，$D(P||\widehat{P_2}) < \infty$ であるということで，$\widehat{P_2}$ が $\widehat{P_1}$ より良い推定量であるとかは断言できない．今回問題にしているスパースな学習データにおける確率パラメータの推定は，まさにそのような場合であるため，P の推定量 $\widehat{P}$ の良さの尺度として，Kullback-Leibler 情報量ではなく，本節で述べた確率点全体での平均二乗誤差の和を採用した．

3.4 係り受け制約を表す規則の適用確率の推定

本節では，係り受け制約を表す生成規則の適用確率の推定に提案推定量を用いた方法について述べる．ただし，3.3 節では，計算の簡単のために，$k(x,y)$ を式 (9) のように定義したが，実際の係り受け制約を表す規則の適用確率の推定では，2つの生成規則が類似していればいるほど，それらの生成規則の適用確率は近いという傾向があると思われるので，

$$\forall x \in D \;\; k(x,x) = w\,,$$
$$\forall xyz \in D$$
$$[d(x,y) < d(x,z) \text{ ならば } k(x,y) > k(x,z)]\,,$$
$$\forall x \in D \;\; \sum_{y \in D} k(x,y) = 1$$

となるように改良している．

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

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

の適用確率は，

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

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．

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

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

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

と定める．ここで，Rules(δ) 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ータ値に対して，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 $(0 < \alpha < 1)$ 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，ほとんどのパラメータ領域で，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

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

の適用確率は，h' と意味の近い h'' に対する規則

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

の適用確率に近い（すなわち，適用確率の値が局所的に滑らか）と考えられ，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．実際の自然言語文法に対する提案手法の評価を行うために，英語の前置詞句の係り先の判定を確率係り受け文脈自由文法に基づいて行い，その際の生成規則の適用確率の推定に提案手法を用いた実験を行った．比較として，生成規則の適用確率の推定に最尤推定法と代表的なスムージング手法である線形補間法（重み係数の推定は削除補間法による）[3] を用いた実験を行った．本章では，これらの実験とその結果

について述べる．

4.1 実験方法

EDR コーパス[4]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 に係るとき V，n_1 に係るとき N）を取り出して

$$\langle cv, cn_1, p, cn_2, Cat\rangle \qquad (22)$$

を収集した（総数8400）．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，

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

とする．ただし，各 $cv \in$ CV，各 $cn, cn_1, cn_2 \in$ CN，各 $p \in$ 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，

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

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

$$\mathrm{VIP}(cv) \longrightarrow \mathrm{VIP}(cv)\,\mathrm{PP}(-cv), \qquad (23)$$

$$\mathrm{NP}(cn_1) \longrightarrow \mathrm{NP}(cn_1)\,\mathrm{PP}(-cn_1), \qquad (24)$$

$$\mathrm{PP}(-cv) \longrightarrow \mathrm{PP}(cn_2,p), \qquad (25)$$

$$\mathrm{PP}(-cn_1) \longrightarrow \mathrm{PP}(cn_2,p) \qquad (26)$$

の適用確率である．したがって，$\langle cv, cn_1, p, cn_2\rangle$ において cn_2 を head 概念，p を function とする前置詞句は，

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

ならば cv に対応する動詞に係り（$Cat = V$），

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

ならば cn_1 に対応する名詞に係る（$Cat = N$）と判定できる．

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

の場合は判定不能である．実験では，式 (22) の形態の5つ組からなる8400個のデータを10等分し，その1つをテストデータとし，残りから学習データ（構文木列）を作成し☆☆，これを用いて適用確率を推定した．テストデータに対して上記の方法で係り先の判定を行い，テストデータを変えてこれを繰り返し，正解率を求めた．

式 (23)，(24) 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式 (25)，(26) の形態の生成規則の適用確率の推定は，最尤推定，3.4 節で述べた提案手法に基づく推定（ただし，$\alpha = w = 0.9$，$R = 2$ とした☆☆☆），および，線形補間法の3つの手法で行った．

今回行った線形補間では，係り受け制約を表す生成規則の適用確率の推定に対して，次のような補間式を考えた．

$$p(X(-h) \longrightarrow X(h',f)) = \sum_{\ell=0}^{3} \lambda_\ell \cdot p_\ell(X(-h) \longrightarrow X(h',f)) \qquad (27)$$

p_1，p_2，p_3 は係り受け規則を表す生成規則の適用確率を与える関数で，それぞれ以下のような制約を満たす．ただし，$\mathrm{Sup}(h)$ は h の直接の上位概念を表す．

(C1) $\mathrm{Sup}(h_1') = \mathrm{Sup}(h_2')$ なる任意の h_1'，h_2' に対して

$$p_1(X(-h) \longrightarrow X(h_1',f)) = p_1(X(-h) \longrightarrow X(h_2',f))$$

(C2) $\mathrm{Sup}(\mathrm{Sup}(h_1')) = \mathrm{Sup}(\mathrm{Sup}(h_2'))$ なる任意の h_1'，h_2' に対して

$$p_2(X(-h) \longrightarrow X(h_1',f)) = p_2(X(-h) \longrightarrow X(h_2',f))$$

(C3) 任意の h_1'，h_2' に対して

$$p_3(X(-h) \longrightarrow X(h_1',f)) = p_3(X(-h) \longrightarrow X(h_2',f))$$

p_0 に対しては何の制約も考えない．また，λ_0，λ_1，λ_2，λ_3 は重み係数で，$\lambda_0+\lambda_1+\lambda_2+\lambda_3 = 1$ であり，削

☆ その単語の直接の上位概念．

☆☆ 前置詞句の係り先が分かっているので，先の文法に基づき一意に構文木が得られる．

☆☆☆ 本来は，$\mathrm{PP}(-c)$（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．

表3 前置詞句の係り先の判定実験結果
Table 3 The result of the experiment in PP-attachment.

	correct		error		indecisive	
	%	(num)	%	(num)	%	(num)
maximum likelihood estimator	6.0	(246)	0.5	(19)	93.5	(3824)
proposed method	58.9	(2407)	15.5	(634)	25.6	(1048)
linear interpolation	58.6	(2398)	15.7	(643)	25.6	(1048)

表4 テストデータを制限した前置詞句の係り先の判定実験結果
Table 4 The result of the experiment in PP-attachment, when the test data is restricted.

F	総数	maximum likelihood estimator				proposed method				linear interpolation			
		correct		error		correct		error		correct		error	
		%	(num)	%	(num)	%	(num)	%	(num)	%	(num)	%	(num)
1	3041	8.1	(246)	0.6	(19)	79.2	(2407)	20.8	(634)	78.9	(2398)	21.1	(643)
2	2062	10.4	(214)	0.8	(16)	80.9	(1669)	19.1	(393)	80.4	(1658)	19.6	(404)
3	1615	11.8	(191)	0.7	(12)	81.4	(1314)	18.6	(301)	80.7	(1304)	19.3	(311)
4	1262	12.7	(160)	0.9	(11)	81.8	(1032)	18.2	(230)	81.0	(1022)	19.0	(240)
6	879	12.3	(108)	0.8	(7)	82.9	(729)	17.1	(150)	82.1	(722)	17.9	(157)
8	646	13.5	(87)	1.1	(7)	83.0	(536)	17.0	(110)	82.5	(533)	17.5	(113)
10	514	14.8	(76)	0.8	(4)	85.6	(440)	14.4	(74)	84.4	(434)	15.6	(80)
15	277	15.9	(44)	0.7	(2)	85.9	(238)	14.1	(39)	84.5	(234)	15.5	(43)
20	143	23.8	(34)	0.7	(1)	87.4	(125)	12.6	(18)	86.0	(123)	14.0	(20)

The test data $\langle cv, cn_1, p, cn_2, Cat\rangle$ satisfies $max\{\mathrm{freq}(\mathrm{PP}(-cv) \rightarrow \mathrm{PP}(*,p)), \mathrm{freq}(\mathrm{PP}(-cn_1) \rightarrow \mathrm{PP}(*,p))\} \geq F$, where $\mathrm{freq}(\delta)$ means the frequency of the production rule δ in the sample trees, and '$*$' means some noun concept.

除補間法により推定した．式(27)による補間は，本来の確率係り受け文脈自由文法，headを単語概念の1つ上の概念でクラスタリングした場合の確率係り受け文脈自由文法，headを単語概念の2つ上の概念でクラスタリングした場合の確率係り受け文脈自由文法，headを区別しない場合の確率係り受け文脈自由文法，これら4つの確率文法（当然，順に粗いモデルとなっている）の適用確率の線形和で係り受け制約を表す生成規則の適用確率を補間するものである．なお，今回使用したEDRのシソーラスは，直接の上位概念が複数存在する概念がある．直接の上位概念が複数存在する名詞概念や，2つ上の上位概念が複数存在する名詞概念が含まれる場合は，式(27)のように単純に線形補間することができない．このため，今回の実験では，式(22)の5つ組はどれも名詞概念 cn_1，cn_2 の直接の上位概念，2つ上の上位概念がそれぞれ1つであるものだけをEDRコーパスから抽出し（総数8400個），これを利用した．

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，$\mathrm{PP}(-cv)$ あるいは $\mathrm{PP}(-cn_1)$ 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，3つの推定法はどれも式(25)あるいは(26)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51.32%存在した．

$\mathrm{PP}(-cv)$ および $\mathrm{PP}(-cn_1)$ 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．判定不能とされるのは，前置詞句が動詞に係る構文木 T_V の生起確率 $P(T_V)$，および，名詞に係る構文木 T_N の生起確率 $P(T_N)$ が等しい場合であるが，今回の実験で判定不能となったのは，すべて，$P(T_V) = P(T_N) = 0$ の場合であった．最尤推定法の場合は，学習データのスパースネスのために，そのような場合が非常に多く，判定不能が93.5%で，正解率はわずかに6%である．一方，提案手法および線形補間法の場合は，スムージングの結果，$P(T_V) = P(T_N) = 0$ となる場合が激減し，正解率はともに60%弱であった．このように，提案手法，線形補間法は学習データのスパースネスに強い確率パラメータの推定法であるといえる．

しかしながら，60%という正解率自体はそれほど高いものではない．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78.0%，誤り率21.7%，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，$\mathrm{PP}(-cv)$ および $\mathrm{PP}(-cn_1)$ 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，提案手法や線形補間法に基づく判定法よりも高い正解率になっていることが分かる．これは，あまりに学習データがスパースであることに起因していると思われる．そこで，ある程度学習データが多くなった場合の傾向を見るために，

$$\mathrm{PP}(-cv) \rightarrow \mathrm{PP}(*,p) \quad (28)$$

の学習データ中での頻度，あるいは，

$$\mathrm{PP}(-cn_1) \rightarrow \mathrm{PP}(*,p) \quad (29)$$

の学習データ中での頻度の高い方が F $(=1\sim20)$ 以上であるテストデータ $\langle cv, cn_1, p, cn_2\rangle$ についてのみ正解率と誤り率を求めたものが**表 4** である．$F=1$ のときに，先の単純なヒューリスティックに基づく手法の正解率と同程度，F が大きくなるに従って，提案手法，線形補間法とも高い正解率となっている．

線形補間法では，生成規則 δ の適用確率を推定する場合に，学習データ中での δ 以外の生成規則の頻度も考慮に入れてスムージングを行っているという点は，提案手法と同じであるが，非常に粗いスムージングとなっている．したがって，ある程度学習データのサイズが大きくなると，線形補間法は提案手法に比べ適用確率の推定が粗くなり，線形補間法に基づく曖昧さ解消法の正解率が提案手法に基づく曖昧さ解消法の正解率より低くなることが予想される．テストデータのサイズを考慮に入れると表3，表4に示した正解率の差は顕著なものではなが，一応予想された結果が得られている．

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，最尤推定によって高信頼度のパラメータ推定値が得られる程度に大きなサイズの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，確率係り受け文脈自由文法に基づいた英語の前置詞句の係り先の判定実験により，提案手法の有効性を示した．

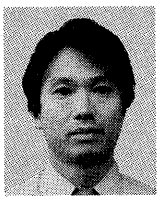
参考文献

1) Fukunaga, K.: *Introduction to Statistical Patten Recognition*, ACADEMIC PRESS INC., Orlando (1972).
2) Hogenhout, W.R. and Matsumoto, Y.: Training Stochastical Grammars on Semantic Categories, *IJCAI'95 Workshop on New Approaches to Learning for Natural Language Processings*, pp.65–70 (1995).
3) 北 研二，中村 哲，永田昌明：音声言語処理，森北出版 (1996).
4) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書 (1995).
5) 白井清昭，乾健太郎，徳永健伸，田中穂積：統計的構文解析における構文的統計情報と語彙的統計情報の統合について，自然言語処理，Vol.5. No.3, pp.85–106 (1998).
6) 田辺利文，冨浦洋一，日高 達：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, pp.95–101 (1995).
7) 田辺利文，冨浦洋一，日高 達：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，Vol.1, No.1, pp.91–94 (1996).
8) 田辺利文，冨浦洋一，日高 達：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，Vol.2, No.1, pp.93–97 (1997).

(平成 10 年 9 月 25 日受付)
(平成 11 年 9 月 2 日採録)

冨浦 洋一（正会員）

昭和 59 年九州大学工学部電子工学科卒業．平成元年同大学院工学研究科電子工学専攻博士課程単位取得退学．同年九州大学工学部助手，平成 7 年同助教授，平成 8 年同大学院システム情報科学研究科助教授，現在に至る．工学博士．自然言語処理，計算言語学，人工知能に関する研究に従事．電子情報通信学会，人工知能学会，言語処理学会各会員．



日高 達（正会員）

昭和 40 年九州大学工学部電子工学科卒業．昭和 44 年同大学院工学研究科電子工学専攻博士課程中退．同年九州大学工学部助手，昭和 48 年同講師，昭和 55 年同助教授，昭和 63 年同教授，平成 8 年同大学院システム情報科学研究科教授，現在に至る．工学博士．形式言語の方程式論，自然言語処理，手書き文字認識の研究に従事．電子情報通信学会，人工知能学会，言語処理学会各会員．