自然言語処理　５６－３
（１９８６．７．２５）

確率文節文法による構文解析

松延　栄治　・　日高　達　・　吉田　将
（九州大学　総合理工学研究科）

我々は、これまで日本語の文節文法の確率文法化の研究を行なってきた。ここではこの文法を確率文節文法と呼ぶ。本稿では先ず確率文節文法の構文解析アルゴリズムについて述べ、次に解析実験とその評価について述べている。解析実験は確率文節文法に基づく解析システムと文節構造解析の優れた方法として知られる文節数最小法に基づくシステムを試作し、平仮名べた書き文１０００文及び漢字仮名混りべた書き文１０００文に対して行ない、２つの方法による結果を比較し、確率文節文法による解析の有効性を確かめた。

"Syntactic Analysis by Stochastic BUNSETSU Grammar" (in Japanese)

Eiji MATSUNOBU , Toru HITAKA and Sho YOSHIDA

Interdisciplinary Graduate School of Engineering Sciences,Kyushu University
6-1 Kasuga-koen,Kasuga-shi,Fukuoka-ken 816 Japan

We have been studying a stochastic grammar which produces strings of BUNSETSU in Japanese. We call this grammar stochastic BUNSETSU grammar. Syntactic analysis method of Japanese sentences by the stochastic BUNSETSU grammar is presented. Then,the result is discussed compared with the LBN(least BUNSETU's number) method which is well-known as a syntactic analysis method for mophological analysis.

１．まえがき

我々は、これまで日本語の文節文法の確率文法化の研究を行なってきた。以下ではこの文法を確率文節文法と呼ぶ。文法の確率化で特に問題になるのは如何にして書き換え規則に適切な確率を付与するかであるが、特に推定困難な終端記号(単語)への書き換えの確率を付与するのに単語の造語モデルを用いた。[1)]

本稿では先ず確率文節文法の構文解析(文節構造解析)アルゴリズムについて述べ、次に解析実験とその評価について述べる。ここで文節構造解析とは入力文に対して文を構成する各単語の切れ目とそれぞれの単語の品詞と活用情報(活用の種類、活用形)等の文法情報を認定することをいう。上記のアルゴリズムに基づく解析システムと文節数最小法[2)]のシステムを試作し、平仮名べた書き文１０００文及び、漢字仮名混りべた書き文１０００文に対して比較実験を行なった。解析結果の平均誤り率で比較して、確率文節文法のほうが平仮名べた書き文で約８．３%、漢字仮名混りべた書き文で約４．３%小さな値を得た。

文節構造解析においては既に優れた解析手法として文節数最小法がある。文節数最小法は漢字仮名混り文のように曖昧さが少ないものにたいしては有効であるが、平仮名べた書き文の解析では同音異字語の問題が解決されず、文節数最小という制約を満たす解は多数存在する。これらの解に尤らしさの順序づけを行なうには他のファクターを持ち込む必要がある。それに対して確率文法は文節構造規則を確率化すると共に単語の生起確率も評価している。複数の解に対して尤らしさの順序づけを確率を用いてほぼ一意に行なうことができる。

この確率文節文法は音声認識、文字認識などのパターン認識の認識率を向上させる後処理にも有効に利用できる。[3)]

２．確率文節文法の構成

日本語の統語規則は単語の連鎖としての文節を規定する規則（文節構造規則）と文節の連鎖としての文を規定する規則（係受け構造規則）の２つで捉えることができる。以下では、確率文節文法を構成する上で基礎となる文節構造規則について述べる。

【定義１】

単語Wの綴りｗ，品詞Ｈ，活用形Ｋの３組（ｗ,Ｈ,Ｋ）をWの単語構造と呼ぶ。但し、本稿では、品詞Ｈはカ行五段活用動詞、サ行変格活用動詞、名詞、形容詞等のように活用の種類が特定できるように通常の品詞を細分類したものである。また、付属語は、一語一品詞として扱う。文は、自立語W_1とそれに続く有限個(≧０)の単語$W_2,W_3,\cdots\cdots,W_n$からなる。W_kの単語構造を(w_k,H_k,K_k)で表すと、文ｂ＝$W_1W_2\cdots W_n$の統語規則は、文中の単語の間の連接可能性を示す３変数述語Ｃにより次のように表される。

$$\bigwedge_{k=0}^{n} C(H_k,K_k,H_{k+1}) \quad \cdots\cdots\cdots \text{[文節構造規則]}$$

但し、$H_0=K_0=H_{n+1}=K_{n+1}=\sqcup$

ここで、⊔は特別に導入された特殊記号で、H_1が自立語の品詞であるときのみ$C(\sqcup,\sqcup,H_1)$＝true、H_nとK_nが文の末尾になり得る品詞，活用形であるときのみ$C(H_n,K_n,\sqcup)$＝trueである。 ■

従来の文節構造規則は、文節内の単語の連接可能性を規定するものであるが、ここでは文内の単語の連接可能性を規定するものに拡張されている。

次に、確率文節文法を確率正規文法の枠組で記述すれば次のようになる。

$G=\langle N,\Sigma,Ps,S\rangle$

$N=\{S\}\cup\bigcup_i\{H_i,(H_i,K_1),(H_i,K_2),\cdots\cdots\}$

$\Sigma=\bigcup_i\{w_i\}$

Ｐｓ:次の２つのタイプがある。

(ｉ) $S\xrightarrow{q}H$ ……(1)

$(H,K)\xrightarrow{r}H'$ ……(2)

$(H,K)\xrightarrow{s}\varepsilon$ ……(3)

(ii) $H\xrightarrow{t}w(H,K)$ ……(4)

ここで(1),(2),(3)はそれぞれ文節構造規則の$C(\sqcup,\sqcup,H)$,$C(H,K,H')$,$C(H,K,\sqcup)$を反映している。タイプ(ii)は品詞から終端記号すなわち具体的な単語Wの綴りｗへの生成規則と考えられる。ここでｑ,ｒ,ｓ,ｔは次式で与えられる。[1)]

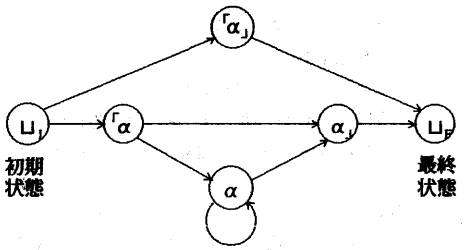
$q=p(H\mid\sqcup)$

$r=p(H'\mid(H,K))$

$s=p(\sqcup\mid(H,K))$

$t=p(K\mid H)\,p(MW\mid H)$

タイプ(ｉ)の生成規則に伴う確率，即ち、ｑ,ｒ,ｓは、非終端記号の数が少ないため、ある程度の規模のテキスト集合より推定可能である。しかし、タイプ(ii)の生成規則は数が多く(数万)、精度良く、確率を推定するには莫大な量のテキストデータを収集する必要がある。しかし、１０万近い一般の単語が各々一回以上出現するテキストデータを収集するのは困難である。$p(K\mid H)$の値はある程度の規模のテキスト集合より推定可能であるの。MWは活用語尾がない場合は単語の綴りそのものであるので、以後混同の恐れのないかぎり$p(MW\mid H)$を単語の生起確率と呼ぶ。単語の生起確率を推定するために単語の造語モデルを構成している。(1)日本語では意味を担う最小単位は漢字で



図－1 単語の造語モデル(品詞別)

あり、漢字が意味的に結合して単語が造語される。(2)造語過程はマルコフ過程である。以上より漢字の結合、即ち、造語過程をラベル付き1重マルコフモデル(図-1)でモデル化している。品詞別に造語モデルを構成し、このモデルから単語の生起確率を求めている。
ここでは漢字を4つのタイプに細分したものを遷移の単位としている。即ち、述語B(α)で漢字αが単語の先頭であることを、E(α)で単語の末尾であることを表わすとすれば、

(1)「α」‥‥ B(α)∧ E(α)
(2)「α ‥‥ B(α)∧~E(α)
(3) α ‥‥ ~B(α)∧~E(α)
(4) α」‥‥ ~B(α)∧ E(α)

の4つである。一般に品詞Hの語幹MWが$\alpha_1\alpha_2\cdots\alpha_n$の単語列からできているとすると、p(MW|H)は次式 により計算できる。

$$p(MW|H)=p(\alpha_1|\sqcup_I)\cdot p(\alpha_2|\alpha_1)\cdots\cdots p(\alpha_n|\alpha_{n-1})\cdot p(\sqcup_F|\alpha_n)$$

但し、$p(\alpha_i|\alpha_{i-1})$は品詞Hにおける遷移確率の値である。例えば、"日本語"という単語は$\sqcup_I$→「日→本→語」→$\sqcup_F$のように状態推移して生成されると考える。この時、遷移確率の積p(「日|$\sqcup_I$)・p(本|「日)・p(語」|本)・p($\sqcup_F$|語」)は名詞における"日本語"という単語の生起確率の値である。

3．構文解析アルゴリズム

以下の説明で必要となる諸定義を述べておく。確率文法における導出は、導出の生起確率を除き一般の形式文法の場合と同様に定義される。[4)]

【定義2】

α，$\beta\in(N\cup\Sigma)^*$、かつ$A\xrightarrow{p}\beta\in P$sであれば、$\alpha A\gamma\overset{p}{\Longrightarrow}\alpha\beta\gamma$と書く。さらに、$\alpha_0,\cdots,\alpha_n\in(N\cup\Sigma)^*$で、かつ$\alpha_{i-1}\overset{p_i}{\Longrightarrow}\alpha_i$(i=1,2‥,n)であれば、$\alpha_0\overset{p_1}{\Longrightarrow}\alpha_1\overset{p_2}{\Longrightarrow}\cdots\overset{p_n}{\Longrightarrow}\alpha_n$を$\alpha_0$から$\alpha_n$の導出と呼び、その生起確率pは$p=p_1p_2\cdots p_n$で与えられる。以後簡略のため、上記の導出を誤解の恐れのない限り、単に$\alpha_0\overset{p}{\Longrightarrow}\alpha_n$と書く。上記の導出の各ステップにおいて、最左の非終端記号に対して確率生成規則が適用されている時、この導出を最左導出といい、$\alpha_0\overset{p}{\underset{lm}{\Longrightarrow}}\alpha_n$と書くが、以後導出とは、最左導出をさしlmを省略する。また$\overset{p}{\underset{*}{\Longrightarrow}}$は0回以上、書き替え規則を適用してえられる最左導出を示す。但し、0回の場合の確率は1とする。$\overset{p}{\Longrightarrow}$は正しく1回の書き換え規則の適用による最左導出を表わすとする。Gにおける最左導出の全体集合をD_Gと書く。Gによって生成される確率言語L(G)は次のように定義される。

$$L(G)=\{(w,p(w))\mid w\in\Sigma^*,\ p(w)=\Sigma p_i,\ S\overset{p_i}{\underset{*}{\Longrightarrow}}w\in D_G\}$$

(w,p(w))∈L(G)のときGにおけるwの生起確率はp(w)であるという。

任意のGに対して、確率生成規則の確率を無視して得られる文法を$\overline{G}$と書く。Dを生成規則の集合P上で定義される確率としてPs=(P,D)と置けば、$\overline{G}=\langle N,\Sigma,P,S\rangle$ となる。L(G)は(L,p)により特徴づけられ、L(G)=(L,p)である。ここでLは言語，pはL上で定義される確率分布である。この時、$L(\overline{G})=L$が成り立つ。 ■

一般に、Gにおける$\alpha\in(N\cup\Sigma)^*$から$\beta\in(N\cup\Sigma)^*$の最左導出は複数個存在しうる。従って、ある入力文字列(文)$s\in\Sigma^*$ が与えられたとき、次の2つの問題が考えられる。

(1) Sからsの最左導出の中で最大の生起確率を持つものを求める
(2) p(s)を求める

(1)の問題は文$s\in\Sigma^*$の文法構造の中で一番確からしいものを求める文法解析の問題であり、(2)は文sが生起する確率を求める問題である。

ここでは(1)の最大の生起確率を持つ文法構造を求める問題に対する解析アルゴリズムを中心に述べ、(2)の問題に関しては、補足的に述べる。

解析アルゴリズムは、解析手法として動的計画法の手法である表方式を用いており、パーズリスト作成アルゴリズムとパーズリストから解を抽出する解抽出アルゴリズムとからなっている。解析の評価値として文法構造の生起確率pを用いて、評価値最大の文法構造を1つ抽出するアルゴリズムである。以下の議論では入力文字列sをn長さの文字列$a_1a_2\cdots a_n$とする。

【定義3】

集合Γ(i) (0≦i≦n-1)を次のように定義する。

$$\Gamma(i)=\{H\xrightarrow{p}w(H,K)\in P\mid a_{i+1}\cdots a_j=w,\ 0\le i<j\le n\}$$

■

集合Γ(i)を求めることは、終端記号列が、入力文字列の部分列$a_{i+1}\cdots a_j$の最左部分語となっているすべてのタイプ(ii)の確率生成規則を求めることに相当

する。

整数ｉ，ｊ、品詞と活用形の組〈H,K〉,実数ｐから成る４項系列[ｉ,ｊ,〈H,K〉,ｐ]を項目といい、ｐを項目の評価値と呼ぶ。パーズリストPLは項目の或る有限集合であり、PLの項目の中で第２要素が整数ｊである全ての項目からなる集合を部分リストPL_jと呼ぶ。

$PL=\bigcup_{i=1}^{n} PL_i$である

《パーズリスト作成アルゴリズム》

入力　：入力文字列$a_1a_2\cdots a_n$

　　：確率文節文法Ｇ＝〈Ｎ,Σ,Ｐs,Ｓ〉

出力　：パーズリストPL

初期設定：$PL_i=\phi\ (i=1,\cdots,n)$, $p_{max}=0$

方法　：

〈ステップ１〉

集合Γ(Ｏ)を求め、その各要素$H\xrightarrow{p}a_1\cdots a_j(H,K)$について$S\xrightarrow{q}H$なる生成規則があれば、項目[Ｏ,ｊ,〈H,K〉,pq]を実引数として登録ルーチンを実行。

ｉを１から１ずつ増してn-1までステップ２を実行。

〈ステップ２〉

PL_iが空集合でないならば集合Γ(ｉ)を求める。

集合Γ(ｉ)の各要素$H\xrightarrow{p}a_{i+1}\cdots a_j(H,K)$についてつぎの処理を行う。

PLiに[ｋ,ｉ,〈H',K'〉,ｑ]なる項目が存在し、かつ$(H',K')\xrightarrow{r}H$なる生成規則があれば、項目[ｉ,ｊ,〈H,K〉,pqr]を実引数として登録ルーチンを実行する。

《登録ルーチン：引数[i, j,〈H,K〉,ｐ]》

ｊ＝ｎの場合

$(H,K)\xrightarrow{q}\varepsilon$なる生成規則があれば、次の処理の(ｉ),(ⅱ),(ⅲ)のいずれかを実行する。

(ｉ) ｐｑ＞p_{max}の場合

PL_nの全ての項目を削除し($PL_n\leftarrow\phi$)、

項目[i, n,〈H,K〉,pq]をPL_jに登録する。

また、$p_{max}\leftarrow$ｐｑにする。

(ⅱ) ｐｑ＝p_{max}の場合

項目[i, n,〈H,K〉,pq]をPL_nに登録する。

(ⅲ) ｐｑ＜p_{max}の場合

登録ルーチンを終了する。

ｊ≠ｎの場合

PL_jに[i',ｊ,〈H,K〉,ｐ']なる項目が登録されているかどうか検索し次の(ｉ)又は(ⅱ)を実行する。

(ｉ) 登録されていない場合

項目[i, j,〈H,K〉,ｐ]をPL_jに登録する。

(ⅱ) 登録されている場合

1)ｐ＞ｐ'ならば、全ての項目[i',ｊ,〈H,K〉,ｐ']をPL_jから削除し、項目[i, j,〈H,K〉,ｐ]をPL_jに登録する。

2)ｐ＝ｐ'かつｉ≠ｉ'ならば、

項目[i, j,〈H,K〉,ｐ]をPL_jに登録する。

3) 1),2)以外ならば、登録ルーチンを終了する。

■

【定理１】

パーズリスト作成アルゴリズムを実行した結果、[i, j,〈H,K〉,ｐ]∈PL_jであることと、以下のこととは同等である。

(ｉ)ｊ≠ｎの場合

$S\overset{q_{kmb}}{\underset{*}{\Longrightarrow}}a_1a_2\cdots a_mH$

$\overset{r_{mbc}}{\Longrightarrow}a_1\cdots a_ma_{m+1}\cdots a_j(H,K)$

∧ $p=max\{q_{kmb}r_{mbc}\}$

∧ ｉは$p=max\{q_{kmb}r_{mbc}\}$を満たすｍである。

(ⅱ)ｊ＝ｎの場合

$S\overset{q_{kmb}}{\underset{*}{\Longrightarrow}}a_1a_2\cdots a_mH_b$

$\overset{r_{mbc}}{\Longrightarrow}a_1\cdots a_ma_{m+1}\cdots a_n(H_b,K_c)$

$\overset{s_{bc}}{=\!\!\Rightarrow}a_1\cdots a_ma_{m+1}\cdots a_n$

∧ $p=max\{q_{kmb}r_{mbc}s_{bc}\}$

∧ ｉは$p=max\{q_{kmb}r_{mbc}s_{bc}\}$を満たすｍであり、更に、この時、$H=H_b$, $K=K_c$である。■

$a_1\cdots a_n\in L(G)$である$a_1\cdots a_n$を生成する最左導出の生起確率ｐが最大であるための必要十分条件は、[i, n,〈H,K〉,ｐ]の形式の項目がPL_nに少なくとも１つ存在することである。

【定理２】

パーズリスト作成アルゴリズムの最大時間計算量、最大領域計算量は入力文字列の長さｎに対して共にＯ(ｎ)である。■

パーズリスト作成アルゴリズムによって作成したパーズリストを参照して入力文字列$a_1a_2\cdots a_n$に対する最左導出を行う確率生成規則の系列を抽出するアルゴリズム(解抽出アルゴリズム)について述べる。

このアルゴリズムは評価値最大(生起確率最大)の最左導出に対する確率生成規則の逆系列を１つだけ抽出する。計算状況Ｒを整数ｉ,品詞の種類Ｈ,実数ｐからなる３項系列(ｉ,Ｈ,ｐ)で表わす。

《解抽出アルゴリズム》

入力：パーズリスト$PL_1,PL_2,\cdots,PL_n$

　　：確率文節文法Ｇ＝〈Ｎ,Σ,Ｐs,Ｓ〉

出力：入力文字列$a_1a_2\cdots a_n$に対する最左導出の

　　の逆系列

方法：

〈ステップ１〉

PL_nの項目[i, n,〈H,K〉,ｐ]∈PL_nを任意に取り出

し、$(H,K)\xrightarrow{q}\varepsilon\in Ps$と$(H,K)\xrightarrow{q}\varepsilon\in Ps$を検索(必ず存在する)し、出力する。さらに、計算状況Rを(i,H,p/qr)とする。もし、$PL_n=\phi$ならば、エラーを出力して解抽出アルゴリズムを終了する。

. 計算状況R(j,H,p)の第一要素jが0になるまでステップ2を繰返す。

〈ステップ2〉

$p=p'q$であるような

$[i,j,\langle H',K'\rangle,p']\in PL_j$

$(H',K')\xrightarrow{q}H\in Ps$

$H'\xrightarrow{r}a_{i+1}\cdots a_j(H',K')\in Ps$

を検索(必ず存在する)し、$(H',K')\xrightarrow{q}H$

$H'\xrightarrow{r}a_{i+1}\cdots a_j(H',K')$を出力する。計算状況Rを(i,H',p/qr)とする。

〈ステップ3〉

ステップ2を終了した時の計算状況Rを(0,H,p)とすると、$S\xrightarrow{p}H\in Ps$を検索(必ず存在する)し、出力する。■

【定理3】

入力文字列$a_1a_2\cdots a_n$が$a_1a_2\cdots a_n\in L(G)$ならば、解抽出アルゴリズムは終了し、$a_1a_2\cdots a_n$に対する最左導出を行う確率生成規則の系列を逆順に出力する。■

【定理4】

解抽出アルゴリズムを1回実行するのに要する最大時間計算量，最大領域計算量は共にO(n)である。■

次に文字列の生起確率を求めるアルゴリズムについてのべる。文字列$a_1a_2\cdots a_n$が$a_1a_2\cdots a_n\in L(G)$であるとき$p(a_1a_2\cdots a_n)$を求めるアルゴリズムは前述したパーズリスト作成アルゴリズムの登録ルーチンと項目の内容を若干変更すればよい。すなわち、[i, j,〈H,K〉,p]の代りに[j,〈H,K〉,p]を項目とする。p_{max}の代わりにp_{sum}を用いる。初期設定で$p_{sum}=0$とする。登録ルーチンは次のように変更する。

《登録ルーチン：引数[j,〈H,K〉,p]》

j＝nの場合

$(H,K)\xrightarrow{q}\varepsilon$なる生成規則があれば、

$p_{sum}\leftarrow p_{sum}+pq$とする。

j≠nの場合

PL_jに[j,〈H,K〉,p']なる項目が登録されているかどうか検索し次の(i)又は(ii)を実行する。

(i) 登録されていない場合

項目[j,〈H,K〉,p]をPL_jに登録する。

(ii) 登録されている場合

$p'\leftarrow p'+p$とする。

この変更したパーズリスト作成アルゴリズムに関して以下の定理が成立する。

【定理5】

パーズリスト作成アルゴリズムを実行した結果、$p_{sum}\neq 0$であるp_{sum}が存在することと、以下のことと同等である。

$$S\xrightarrow[*]{p_i}a_1a_2\cdots a_n\ \wedge\ p_{sum}=\sum_i p_i$$ ■

すなわち、$p_{sum}=0$ならば、$a_1a_2\cdots a_n\notin L(G)$であり、$p_{sum}\neq 0$ならば、$a_1a_2\cdots a_n\in L(G)$であり、$p(a_1a_2\cdots a_n)=p_{sum}$ということである。

4．書き換え規則の確率の付与と実現

タイプ(i)の書き換え規則すなわち非終端記号への書き換えの規則は次のようにして求めている。実際の文1000文程度を予め解析を行なって、それを基にそれぞれの規則に3段階のある値の何れかを与える。次に規則の左辺が同じものについて値の和が1になるように規格化する。タイプ(i)の確率生成規則をPs_iとすると$|Ps_1|$は$|Ps_2|$に比較してかなり小さい。また、$\pi\in Ps_1$は同じ長さ$|\pi|$を持っている。このため、タイプ(i)の確率生成規則$(H,K)\xrightarrow{p}H'$は(H,K)，H'をキーとして確率pが検索できるテーブル(確率接続テーブル) で実現できる。現在、(H,K)の組が605個、Hが205個である。

次にタイプ(ii)の書き換え規則の確率は2．で述べたように品詞別の単語の造語モデルをラベル付き1重マルコフモデルで構成し、このモデルより求める。但し、造語モデルの遷移確率は一般の機械国語辞書[5]の漢字見出しから求めている。$\pi\in Ps_2$の終端記号列(単語の綴り)の長さは一定ではない。また、入力文字列$a_i\cdots a_n(1\leq i\leq n)$の最左部分語が終端記号列となる確率生成規則を能率良く検索できるデータ構造が必要である。このようなタイプ(ii)の確率生成規則は、終端記号列を見出しとする拡張B-tree構造[6]の確率付き辞書で実現できる。現在、終端記号列が自立語のものが約83000個、付属語，形式名詞，補助用言であるものが約600個ある。

5．実験と考察

確率文節文法と文節数最小法による文節構造解析の比較実験とその考察について述べる。解析プログラムはPL/Iで約1000行である。九大大型計算機センターのFACOM M-380S上で実行した。

今回試作のシステムは数詞、固有名詞、接辞の処理が未だ組み込まれていないため、実験の入力文としてはこれらの単語を含まないものが望ましい。そこで武者小路実篤の"人生論"[7] 冒頭部分から1000文を抜き出して、多少手を入れたものを使用した。入力文としては2つを用意した。

(1)漢字仮名混りべた書き文(図－２)

出来るだけ漢字表記したもの

文の平均長３５文字、最大長１８９文字

(2)平仮名べた書き文(図－３)

全て平仮名にしたもの

文の平均長４３文字、最大長２３６文字

表１に文字長別の入力文の個数を示す。

1 自分というものが意識に浮かんだときは既に自分が生まれていたときである。
2 つまりそのとき自分が人間として生まれていたのだ。
3 そして恐らく自分は何年か先に意識を失うであろう。
4 そのとき自分は人間でなくなっているときだ。
5 自分は人間以外の世界をまるで知らない。

図－２　漢字仮名混りべた書き文

1 じぶんというものがいしきにうかんだときはすでにじぶんがうまれていたときである。
2 つまりそのときじぶんがにんげんとしてうまれていたのだ。
3 そしておそらくじぶんはなんねんかさきにいしきをうしなうであろう。
4 そのときじぶんはにんげんではなくなっているときだ。
5 じぶんはにんげんいがいのせかいをまるでしらない。

図－３　平仮名べた書き文

表１　入力文の個数分布

	漢字仮名混りべた書き文	平仮名べた書き文
1~ 10	２６	１１
11~ 20	２３４	１４９
21~ 30	２９０	２１７
31~ 40	１５５	２１５
41~ 50	１１６	１１７
51~ 60	６８	９１
61~ 70	４３	６９
71~ 80	２５	３９
81~ 90	１９	３２
91~100	９	１８
101~110	２	１２
111~120	４	１２
121~130	２	５
131~140	３	２
141~150	２	３
151~160	１	３
161~170	０	１
171~180	０	２
181~190	１	０
191~200	０	０
201~210	０	１
211~220	０	０
221~230	０	０
231~240	０	１

実験１

先ず漢字仮名混りべた書き文に対する実験について述べる。確率文節文法では確率が最大になる解析だけを、文節数最小法では文節数が最小となる解析だけを出力している。

確率文法と文節数最小法との解析の曖昧さの数を比較をする。ここでは、次の２通りの曖昧さを比較する。

(1)構造の曖昧さ

入力文字列を単語の切れ目と各単語の品詞H、活用形Kについてのみ着目した場合の曖昧さである。

(2)全解析の曖昧さ

構造の曖昧さに更に漢字表記の綴りまで区別したものであり、例えば、"うむ"に対して"生む","産む","倦む","膿む","績む"を区別する。

表２に構造の曖昧さの比較を、表３に全解析の曖昧さの比較を示す。

表２　構造の曖昧さの比較

曖昧さの個数	確率文法	文節数最小法
1	* 1000	98
2	0	162
3	0	28
4	0	122
5	0	2
6~10	0	152
11~100	0	316
101~500	0	55
501~1000	0	18
1001~	0	* 47
* 曖昧さの個数の最大値	1	5474304

表３　全解析の曖昧さの比較

曖昧さの個数	確率文法	文節数最小法
1	971	91
2~5	* 29	251
6~10	0	117
11~50	0	270
51~100	0	56
101~1000	0	128
1001~10000	0	49
10001~100000	0	27
100001~1000000	0	7
1000001~	0	* 4
* 曖昧さの個数の最大値	4	1564738560

ある文の正しい解析が全解析の複数の解のなかに存在する場合、ある文の正解が解に含まれるという。１００文中で何文が正解が解に含まれるかを表４に示す。

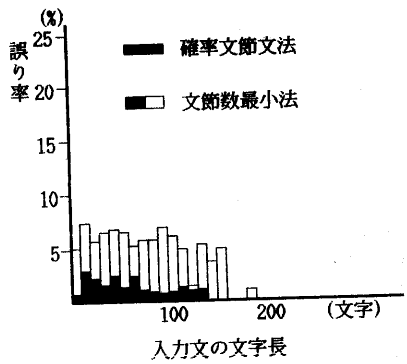
表4　正解が含まれる文の数(1000文中)

確率文法	７９０(文)
文節数最小法	９９９

次に解析結果の品質を評価する。全解析の曖昧さによって複数の解が出力され、本来なら複数の解の平均をとるべきであるが簡単にはとれないため、ここでは１番目に出力される解析結果の誤りを比較する。

$$誤り率 = \frac{ある文の中で解析を誤った部分の長さ}{ある文の入力文字列の長さ}$$

文節最小法、確率文法それぞれの入力文字列の長さに対する誤り率のグラフを図－４に示す。誤り率の平均値の比較を表５に示す。



図－４　誤り率の比較

表５　誤り率の平均の比較

確率文法	２．２(%)
文節数最小法	６．５

実験２

平仮名べた書き文に対する実験について述べる。表６に構造の曖昧さの比較、表７に全解析の曖昧さの比較を示す。表８に正解が解析の中に存在する文の数を示す。図－５に確率文節文法による平仮名べた書き文の解析例を示す。

表６　構造の曖昧さの比較

曖昧さの個数	確率文法	文節数最小法
1	* 1000	42
2	0	70
3	0	24
4	0	79
5	0	5
6~10	0	134
11~100	0	337
101~500	0	132
501~1000	0	35
1001~	0	* 142
* 曖昧さの個数の最大値	1	12541132800

```
  文番号 =     96
じぶんはそれをしりたいとおもうがまだしることはできないのだ。
  文字長 =   30
解析所要時間      590 MSEC   ＩＴＥＭ数    266 <=      290 個
--------------------------------------------------------------------
じぶん  は  それ  を  しり  たい  と  おもう  が  まだ  しる  こと  は  でき
ない  の  だ  。
生起確率   4.54179E-0049
```

じぶん	自分	名詞		10: 10
は	は	係助詞		380:380
それ	其れ	名詞		10: 10
を	を	格助詞		302:302
しり	知り	ラ行五段動詞	連用	130:128
たい	たい	希望の助動詞	終止	634:634
と	と	格助詞		305:305
おもう	思う	アワ行五段動詞	終止	140:140
が	が	接続助詞		340:340
まだ	未だ	副詞		20: 20
しる	知る	ラ行五段動詞	連体	130:131
こと	こと	形式名詞１		269:269
は	は	係助詞		380:380
でき	出来	上一段動詞	未然	150:148
ない	ない	打消の助動詞	連体	579:580
の	の	準体助詞		510:510
だ	だ	断定の助動詞	終止	613:613
。	。	句点		910:910

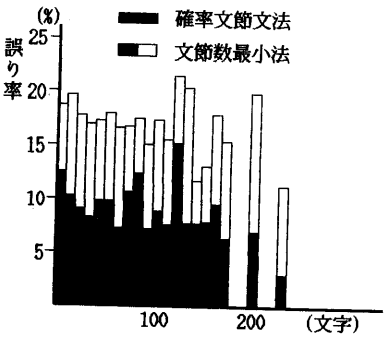
図－５　確率文法による解析例

表７　全解析の曖昧さの比較

曖昧さの個数	確率文法	文節数最小法
1	845	18
2~5	138	53
6~10	11	53
11~50	5	168
51~100	0	65
101~1000	* 1	226
1001~10000	0	142
10001~100000	0	95
100001~1000000	0	65
1000001~	0	* 115
* 曖昧さの個数の最大値	125	$5.2660718*10^{16}$

表８　正解が含まれる文の数(1000文中)

確率文法	３７９(文)
文節数最小法	９６３



図－６　誤り率の比較

表９　誤り率の平均の比較

確率文法	９．３(%)
文節数最小法	１７．６

文節最小法、確率文法それぞれの入力文字列の長さに対する誤り率のグラフを図－６に示す。誤り率の平均値の比較を表９に示す。

平仮名べた書き文で２つの解析の曖昧さを比較すると漢字仮名混りべた書き文に比べて一段と解の絞りこみ方の差が現われてくる。また、文節数最小法は各々の単語の生起に関しては何も評価しておらず、仮名漢変換等においては同音異字の問題が解決されなければならないことを示している。

６．あとがき

確率文節文法による構文解析とその実験結果について述べた。文節数最小法との比較において尤らしい解の絞りこみの違いを示し、更に誤り率の比較のおいて確率文法による解析の有効性を示した。

今回の実験では単語の読みに曖昧さはないとしていたが実際には存在する。これに対処するためには単語は漢字の綴りと読みを持っているとして単語の造語モデルを捉え直し、生起確率を付与することが考えられる。

確率文節文法は確率という概念で全体が統一されているため、解の絞りこみにおいて文節数最小法のように他のファクター即ち単語の頻度情報等を持ち込む必要がない。このためパターン認識で文脈情報を用いて認識率を向上させるような後処理において確率文法を埋め込んでもマッチングが良いことが期待でき、このことを実証することも今後の課題として残っている。

また現在は数詞、接辞の処理が組み込まれていない。これらのモデルを作り、確率文法に組み込む必要がある。

参考文献

1) 松延,日高,吉田: "日本語確率文法における書き換え規則の確率の推定について"情報処理自然言語研究会研資 55-4,1986
2) 吉村,日高,吉田: "文節数最小法を用いたべた書き日本語文の形態素解析"情報処理論文誌 Vol.24, No.1,1983
3) 長田,牧野,日高: "日本語の文脈情報を用いた文字認識"信学論文誌 Vol.J67-D,No.4,1984
4) FU,K.S.: "Sntactic Pattern Recogonition and Applications"Prentice-Hall,1982
5) 吉田,日高,稲永,田中,吉村: "公用データベース日本語単語辞書の使用について"九大大型計算機センター広報,Vol.16,No.4,1983
6) 日高,吉田,稲永: "拡張B-treeによる日本語単語辞書の作成"情報処理自然言語研究会研資 33-8,1986
7) 武者小路実篤: "人生論" 角川文庫,1985