自然言語処理 108－15
(1995. 7. 21)

係り受け制約を含む文脈自由文法

田辺 利文，富浦 洋一，日高 達
九州大学工学部情報工学科
〒812 福岡市東区箱崎6丁目10番1号 092-641-1101 (ex 5372)
E-mail tanabe@lang.ai.kyushu-u.ac.jp
tom@lang.ai.kyushu-u.ac.jp
hitaka@lang.ai.kyushu-u.ac.jp

あらまし 自然言語処理において，入力文に対応する構文構造がたくさん存在することが問題点の一つである．意味的に正しい構文構造が選択できれば処理の質を向上させることが出来る．従来はこれを選択制約で行なっていた．しかし，この制約で用いられる意味分類は非常に粗く，精度が悪いという問題があった．また，確率化を考えた場合，構文構造の確率と係り受け制約をどの程度満足しているかの整合性の問題もあった．本研究報告では，非終端記号を，それから導出される句の概念（意味）によりを細分化し，さらにシソーラスの概念間の上位下位関係を文法規則として捉えることにより，係り受け制約を生成規則中に取り込んだ文脈自由文法を提案する．

キーワード 係り受け制約，ヘッドフレーズ，ヘッドワード，シソーラス，上位下位関係，確率文脈自由文法

A Context Free Grammar Expressing the Dependency Constraint.

Toshifumi Tanabe, Yoichi Tomiura, Toru Hitaka
Department of Computer Science and Communication Engineering,
Faculty of Engineering, Kyushu University.
Hakozaki 6-10-1, Higashi, Fukuoka City, Fukuoka 812, Japan
Tel +81-92-641-1101 (ex 5372)
E-mail tanabe@lang.ai.kyushu-u.ac.jp
tom@lang.ai.kyushu-u.ac.jp
hitaka@lang.ai.kyushu-u.ac.jp

Abstract In the natural language processing by computer it is one of the problems that there are lots of syntactic structures corresponding to the input sentence.If syntactic structures can be chosen correctly in a view of meanings, the quality of processing will be more improved. Up to now, Selectional Restriction has been used conventionally. But the semantic category used by this restriction was very rough and not precise. Besides it was uncertain how much it satisfied probability of the syntactic structure and the dependency constraint.In this report, we propose Context Free Grammers expressing the dependency constraint by production rules. In these grammars nonterminal symbols are subdivided based on what are expressed by phrases derivated from nonterminal symbols, and superordinate-subordinate relations are treated as production rules .

key words the dependency constraint, head phrase, head word, thesaurus, superordinate-subordinate relation, Probabilistic Context Free Grammar

1 はじめに

自然言語処理における構文解析では，一般に入力文に対応する構文構造がたくさん存在し，それらからどのようにして構文構造を選択するかが問題点の一つである．構文構造の中には誤った意味のものも含まれる．意味的に正しい構文構造を選択できるかどうかが，仮名漢字変換や機械翻訳などの以後の処理の質を大きく左右する．従来は文節数最少法のような粗い経験則によって構文構造の間に優先順位を付け，優先順位の高い構文構造に基づき出力を合成していた．しかし，この方法は質の高いものではなかった．そこで，さらに質を上げるには，意味処理の導入が自然言語の機械処理の大きな課題になっている．

しかし，意味処理の徹底した導入は，処理時間の爆発的な増加をもたらし，実用的ではない．従って処理の質と処理時間を考慮した意味処理の一部導入が必要である．意味処理の実用的な導入として，係り受け制約をCFG（文脈自由文法）に組み込むことが考えられる[7]．

また，言語処理に事例データを反映する方法として，文法を確率化することが考えられる．音声認識や文字認識などの要素認識の段階においてはPCFG（確率文脈自由文法）を用いている．そのため，前段に対応して後段である言語認識の段階（構文処理）においても，PCFGを使った方が整合性が良い．

係り受け関係を記述することができ，確率化が容易な文法として，TAG（木接合文法）などが考えられたが[9][10]，強力な構文解析法はまだ開発されておらず，機械処理上での重大な問題点であった．また，CFGに対しては係り受け制約を記述することが難しいとされていたが，出来ないことが証明されたわけではなかった．

本論文では，非終端記号から導出される句の概念（意味）により非終端記号を細分化し，さらにシソーラスを文法規則として捉えることにより，係り受け制約を生成規則中に取り込んだ文脈自由文法を提案する．この文法では，語と語の係り受け関係も生成規則として捉えられるので，構文構造に与えられる確率と選択制約の充足度の整合性を考慮することなく確率文法化することができる．

2 係り受けの形式的定義

2.1 構造的な係り受け関係

句はいくつかの句から構成される．この中で，この句の全体の意味を代表する句が必ず一つ存在することを前提とする．このとき，句の意味を代表する句をその句の *head phrase* と定義する．

$$X \longrightarrow Y \quad \dot{Z}$$

上の記法では，Xの句の中の *head phrase* は Zであることを陽に示している．

Xを root node に持つ部分木において，Xの *head word* を次のように定義する．部分木を構成する生成規則 (*head phrase* を陽に示した) において，Xの *head phrase* がαである場合

- αが終端記号の時，α
- αが非終端記号の時，αを root node とする部分木の *head word*

が X の *head word* である．*head word* は，部分木の中でその句の意味を代表する語になる．

今後，句 Xを，非終端記号 Xの意味として用いる他に，Xから導出されている単語列の意味としても用いることにする．

$$X \longrightarrow Y_1 \quad Y_2 \quad \cdots \quad \dot{Z} \quad \cdots \quad Y_n$$

という規則があり，Y_iの *head word* が w_i，Zの *head word* が wであるとき，w_iは wに構造的に係っている (構造的な係り受け関係) と定義する．この係り受け関係は，文に対して一意に決まる．

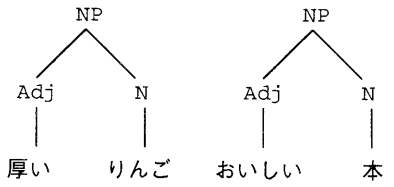
しかし，このCFGは，係り受けに交差を有する文の解析は出来ない．そのような文は少ないので，この類の文は始めから取り扱わないものとする．また，係りの一意性を満足するという仮定のもとで *head phrase* を定義しているので，係り先が二つ以上になる文も除外して扱う．

2.2 意味的に適格な係り受け関係の表現法

構造的な係り受け関係が意味的に適格であるとは限らない．次のような文法の例について考える．

$$
\begin{aligned}
NP &\longrightarrow Adj\ \dot{N} \\
Adj &\longrightarrow \text{厚い} \\
Adj &\longrightarrow \text{おいしい} \\
N &\longrightarrow \text{本} \\
N &\longrightarrow \text{りんご}
\end{aligned}
$$

この生成規則で解析される文(ここでは名詞句)の集合は，意味的には適格であるとは限らない．すなわち，"おいしい本"，"厚いりんご"は，意味的には適格ではないが，2.1節で提案したCFGではこのような文の解析をも許してしまう．



意味的に適格でない場合の例

上の例で「厚い」は「りんご」に構造的に係っているが，意味的には適格に係っていない．そこで意味的にも適格な文のみを解析出来るようにCFGを拡張する．

$$X \longrightarrow Y_1\ \ Y_2\ \ \cdots\ \ \dot{Z}\ \ \cdots\ \ Y_n$$

上の生成規則において，Zから導出される語 w で，Y_iから導出される語 w_iをコントロールすることが出来れば，意味的に適格な係り受け関係も記述することが出来る．従って，

$$X(w) \longrightarrow Y_1(-w)\cdots Z(w)\cdots Y_n(-w)$$

となる．このとき，生成規則のパターンは次の三通りとなる．

$$
\begin{aligned}
X(w) &\longrightarrow Y_1(-w)\cdots Z(w)\cdots Y_n(-w) \\
Y(-w) &\longrightarrow Y(w') \\
X(w) &\longrightarrow w
\end{aligned}
$$

ここで，次のように記法の意味を定義する：

w: *head word* (この場合単語となる)

$X(w)$: *head word* が wであるカテゴリ Xの句(を導出する非終端記号)

$Y(-w)$: *head word* が wである句に意味的に係るカテゴリ Yの句(を導出する非終端記号)

一番目のタイプの生成規則は，*head phrase* Z から導出される *head word*(単語) w によって，Y_i から導出される句の *head word*(単語) w_i (w に係る)がコントロールされることを表し，二番目のタイプの生成規則によって，w' が w に係る係り受けが意味的に適格であることを表している．

以前，CFGでは選択制約を意味処理の過程で行なっていた．選択制約とは，簡単にいえば，意味的に適格な係り受け関係があるかというのを調べ，適格ではないものを排除することである．上の記法では，この選択制約をも生成規則として記述出来ることを意味する．従って，選択制約をも考慮した構文木の生起確率を簡単に決定することが出来る．例えば，一つの文に対する構文木の順位つけを行なう場合，PCFGによる構文木の確率が高いが意味的に適格な係り受け関係があまり成立していないものと，構文木の確率は低いが意味的に適格な係り受け関係が成立している場合，このどちらの構文木の順位を上げるかが問題になってくる．上記の記法では，そのことを全く意識しなくて済む．

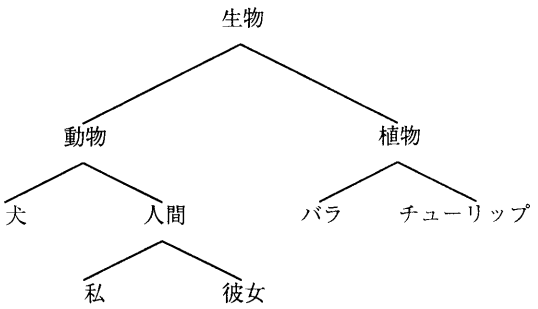
3 生成規則へのシソーラスの組込み

2.2節で提案した文法における二番目のタイプの形の生成規則は，二つの語の間に意味的に適格な係り受け関係があることを意味している．その文法では，*head word*を語(単語)としていた．しかし，この場合では，非終端記号の数が膨大になり，その結果生成規則の数が増え，処理時間も増えてしまう．そのため，意味的に適格な係り受け関係の表現精度にあまり影響を及ぼさないように非終端記号を減らす方法を考える必要がある．

3.1 シソーラス

単語Aが指示する対象の性質を単語Bの指示する対象も持つ時，単語AとBには上位下位関係が存在し，単語AはBの上位語であるという．

シソーラスは概念間の上位下位関係を体系的に表したものである．下にその例を示す．



シソーラスの例

上位下位関係では推移律が成り立つ．例えば,「生物」が「動物」の上位語で，かつ「動物」が「人間」の上位語であるということから,「生物」が「人間」の上位語であることが分かる．

3.2 生成規則への組込み

単語w_1とw_2の間に意味的に適格な係り受け関係が成立していれば，w_1の下位語であるw_1'と，w_2の下位語であるw_2'の間にも意味的に適格な係り受け関係が成立するものと仮定する．

シソーラスにおいて概念W_uとW_dが上位-下位関係であるとき，

$$W_u \longrightarrow W_d$$

の生成規則として捉え，単語wの概念がWであるとき，

$$W \longrightarrow w$$

の生成規則として捉える．

この場合，シソーラスを用いた場合の文法の生成規則のパターンは次の五通りとなる．

$$\begin{aligned}
X(W) &\longrightarrow Y_1(-W)\cdots Z(W)\cdots Y_n(-W)\\
Y(-W) &\longrightarrow Y(W')\\
Y(W) &\longrightarrow W\\
W &\longrightarrow W'\\
W &\longrightarrow w
\end{aligned}$$

但し，記号Wはシソーラス中のある概念記号であり，これを*head word*として用いている．wは単語である．三番目のタイプの生成規則は，非終端記号からシソーラスの概念記号への書換えを行なっている．四番目はシソーラスの概念記号をたどっている．五番目はシソーラスの概念記号から単語への書換えを行なっている．

この時，シソーラスを用いない従来の文法をG_1，シソーラスを代わりに用いた場合の文法をG_2とすると，*head word*を適当に選ぶことにより，

$$L(G_1) = L(G_2)$$

とすることが出来る．シソーラスの比較的上位の概念記号を*head word*とすると，生成規則は減るが意味的に適格な係り受け制約が粗くなる．逆にシソーラスの下位の概念記号を*head word*とすると，意味的に適格な係り受け関係は細かくなるが生成規則の数が膨大になる．そのため，これらを考慮した*head word*の選定が問題となる．

また，文法学習の際には，入力サンプルが多いほどよい．しかし，十分な数の学習サンプルを得ることは容易なことではない．学習の際には限られた入力サンプルで，いろいろな文に対して構文解析出来るようにすることが望まれる．そのためにも，3.2節で述べた方法を用いれば実現できる．

従って，*head word*として，語の代わりにシソーラス上のある概念記号Wを使用すると，ある意味で構文解析能力が拡張される．

4 学習

4.1 確率文脈自由文法

確率文脈自由文法の定義と性質，パラメタ推定法について解説する．確率文脈自由文法PCFGは次のような5組で定義される．

$$G = (\Sigma, V, P, S, p)$$

- Σ : 終端記号の有限集合
- V : 非終端記号の有限集合
- S : 開始記号 $(S \in V)$
- P : $(X \to \alpha | X \in V, \alpha \in (\Sigma \vee V)*)$ の有限部分集合
- p : $P \to (0,1]$ (P から (0,1] への写像)

P の要素は， CFG における書き換え規則である．$X \to \alpha \in P$ と値 $p(X \to \alpha)$ の組を

$$X \overset{p(X \to \alpha)}{\longrightarrow} \alpha$$

で表し，(確率付き) 書き換え規則という．また X,α,$p(X \to \alpha)$ を上の書き換え規則のそれぞれ左辺，右辺， 適用確率という．左辺が同一の非終端記号 (X とする) であるすべての書き換え規則を

$$X \xrightarrow{p_i} \alpha_i \qquad (\text{i=1,2,}\ldots,I_X)$$

とすると， 適用確率の総和は 1 である．

$$p_1 + p_2 + \cdots + p_{I_X} = 1$$

PCFG では文 $s \in L(G)$,s の導出木 T に対し，それぞれ， s， T の生起確率 $P_r(s), P_r(T)$ が次のように定義される．すなわち，$P_r(s)$ は， 文 s を生成するすべての導出木の生成確率の総和であり， $P_r(T)$ は， T の導出に適用された書き換え規則の適用確率の (重複を含めた) 積である．

4.2 パラメタ推定 (最尤推定法)

PCFG $G = (\Sigma, V, P, S, p)$ において， $p : P \to (0,1]$ は標本 (事例データ) に依存して定められるパラメタである．ここでは，パラメタ推定法のうちの最尤推定法について述べる．

N 個の標本 (構文木) を $T_1, T_2, \cdots, T_N$ とし書き換え規則 $X \to \alpha \in P$が構文木 T の導出に適用された回数を $n(T, X \to \alpha)$ で表す．

標本の採集が互いに独立に行なわれたと仮定すると $T_1, T_2, \cdots, T_N$ が標本として収集される確率 (尤度) は，

$$\prod_{k=1}^{N} P_r(T_k)$$

であり，これを最大にする $p(X \to \alpha_i)$ の値は次の式で与えられる．

$$p(X \to \alpha_i) = \frac{\sum_{k=1}^{N} n(T_k, X \to \alpha_i)}{\sum_{i=1}^{I_X} \sum_{k=1}^{N} n(T_k, X \to \alpha_i)}$$

4.3 学習データの作成

3.2節で述べた文法を確率文法化する．生成規則の適用確率の学習は基本的には最尤推定法で行なう．このためには，学習データとして，3.2節で述べた文法による構文木の集合が必要になる．本節では，EDR 電子化辞書のコーパス，概念体系辞書，単語辞書を用いて，学習データを作る方法，および，適用確率の推定法を述べる．なお，本節で扱う文法は，名詞が「の」で連結した名詞句を解析するための文法とする．

EDR コーパスから，複数の名詞が「の」で連結された名詞句と，個々の名詞の概念 (語義) およびその係り受けを抽出することができる．例えば,「円高は当面の景気の足を引っ張る。」に対する，形態素データ，構文木データ，概念関係データから，名詞句「当面の景気の足」，この名詞句における 「当面」の概念記号が 0f32da,「景気」の概念記号が 0eefc9,「足」の概念記号が 0e2f61 であること，および,「当面」が「景気」に係り,「景気」が「足」に係ることが抽出できる．

しかし，学習に必要な構文木は，前節で述べた文法による構文木である．そこで，単語 w_1 の概念記号が C_1 で，単語 w_2 の概念記号が C_2 であり,「w_1 の w_2 」において w_1 が w_2 に係っている場合，以下の 2 つの手順でそれぞれの構文木を求め，2 つを比較する．

(手法 1)

シソーラスにおける C_1 の m 段上位の概念を C_1' ，シソーラスにおける C_2 の m 段上位の概念を C_2' として，以下の最左導出に対応する構文木を求める．

$$
\begin{aligned}
NP(C_2') &\Longrightarrow NP(-C_2')\ \text{の}\ NP(C_2') \\
&\Longrightarrow NP(C_1')\ \text{の}\ NP(C_2') \\
&\Longrightarrow C_1'\text{の}\ NP(C_2') \\
&\overset{*}{\Longrightarrow} C_1\text{の}\ NP(C_2') \\
&\Longrightarrow w_1\text{の}\ NP(C_2') \\
&\Longrightarrow w_1\text{の}\ C_2' \\
&\overset{*}{\Longrightarrow} w_1\text{の}\ C_2 \\
&\Longrightarrow w_1\text{の}\ w_2
\end{aligned}
$$

この場合，m 段遡り終る前に root node に達した場合，root node の概念 C を C_1' 又は C_2' として用いる．

(手法 2)

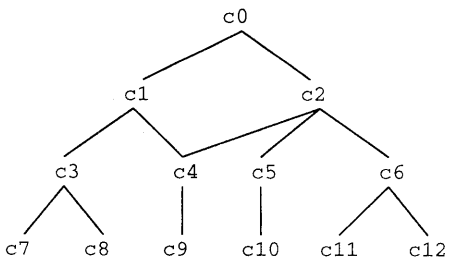
シソーラスにおける C_1 の上位概念のうち，root node からの深さが m の概念を C_1'，シソーラスにおける C_2 の上位概念のうち，root node からの深さが m の概念を C_2' とすると，手法 1 と同様の構文木を求める．

この場合，w_1 または w_2 の概念が root node から深さ m より上位であった場合，その w_1 または w_2 の概念を C_1' 又は C_2' として用いる．

ただし，一つの概念の上位概念が必ずしも一つではないので，各手法において，構文木が複数存在する場合がある．そのような場合は，構文木の数を N とすると，各構文木の頻度を $1/N$ とする．「w_1 の w_2 の w_3」の場合も上記の手法の単純な拡張で，構文木とその頻度を求めることが出来る．

(例)

下の図のようなシソーラスを仮定する．さらに，w_1 に対する概念が C_8，w_2 に対する概念が C_9 と C_{10}，w_3 に対する概念が C_6 であるような単語辞書を仮定する．そして，コーパスから，



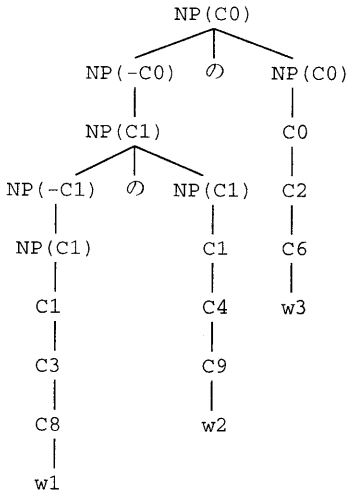
仮定したシソーラス

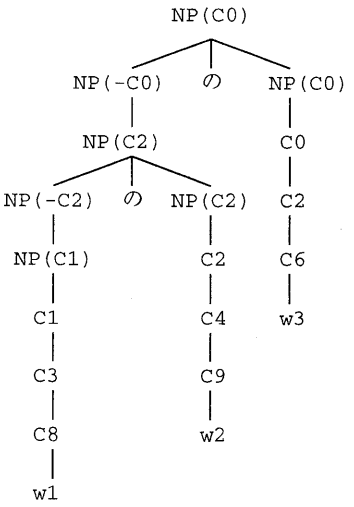
名詞句　w_1の w_2の w_3

各単語の概念　$w_1 - C_8, w_2 - C_9, w_3 - C_6$

係り受け関係　w_1が w_2に係り，w_2が w_3に係る

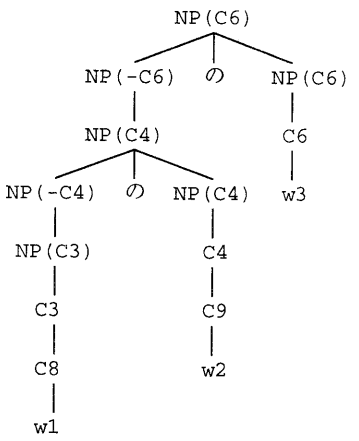
という情報が得られたとする．このとき，手法 1 で m=2 の場合，以下の二つの構文木を作成する．





作成した構文木 (手法 1)

構文木が二つなので，各構文木に与えられる頻度は 1／2 である．また，手法 2 で m=2 の場合，以下の構文木を作成する．



作成した構文木(手法2)

この構文木に与えられる頻度は，1である．

5 おわりに

今後の研究課題については，シソーラスから適当に *head word* を選んで認識実験を行ない，適当な *head word* を正解率，実行時間なども含めて検討する．また，並列の文（*head phrase* が複数出るような場合）も扱えるような文法の提案（一般化），そして認識実験も行なう．

参考文献

[1] 津留正二郎，文節における係り受け情報の抽出，九州大学大学院工学研究科修士論文，昭和５６年２月

[2] 松田和生，日本語文の構文解析九州大学大学院工学研究科修士論文，昭和５７年２月

[3] 元村光男，日本語文の構文解析（係り受けリスト作成アルゴリズム），九州大学大学院工学研究科修士論文，昭和５７年２月

[4] 溝口英樹，名詞シソーラスからの否定的関係の抽出，九州大学大学院工学研究科修士論文，平成２年２月

[5] 苑田良博，大規模コーパスを用いた確率文節文法の精密化，九州大学大学院工学研究科修士論文，平成４年２月

[6] 森重樹，『N1 の N2』の意味構造の推定法，九州大学大学院工学研究科修士論文，平成５年２月

[7] 田辺利文，冨浦洋一，日高達，係り受け関係の記述能力をもつPCFG，電気関係学会第４７回九州支部連合大会講演論文集，pp685，平成６年９月

[8] 日高達，確率文法，情報処理学会誌，Vol.36，No.2，pp.169-176 平成７年２月

[9] Yves Schabes and Richard C.Waters, Lexicalized Context-Free Grammars, In proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, June 1993

[10] Joshi,Aravind K. and Yves Schabes, Tree Adjoining grammars and lexicalized grammars, In Maurice Nivat and Andreas Podellski, editors, Tree Automata and Languages. Elsevier Science 1992