係り受け関係の階層化に基づいた構文木モデルによる構文解析手法の提案

A Syntax Tree Model based on An Extended Japanese Dependency

大野 一樹 † 波多野 賢治 ‡
Kazuki Ohno Kenji Hatano

1. はじめに

日本語は格文法を基本文法として成立しており，日本語における構文解析は文節間の意味的役割を表現する係り受け関係の取得にであると言うことができる．そのため，日本語の既存の構文解析手法はある文節間について係り受け関係が成立しているか否かを判断し，そうして生成された格文法に基づく構文木モデルを利用することによって構文解析を行っている [1, 2].

一方で，英語の構文解析は, 近年，注目されている言語モデルの一つである木置換文法 (TSG: Tree Substitution Grammar) [3, 4] を利用することにより，現在する英語の構文解析手法の中でも最高精度の構文解析精度を実現している [5, 6]．しかし， TSG の構文木生成モデルから生成される構文木は，統語規則から表現される文の統語構造を木構造で表現するのみである．また，格文法を基本文法とする中国語や日本語等の言語では文節間の係り受け関係によって文が構成されるために，語順が自由であるという特徴があり，それゆえ，統語構造を単純に木構造で表現できない．そのため，文節間の係り受け関係が文を構成する文法要素として重要な役割を担っている格文法を基本文法とする中国語や日本語等の言語に対しては木置換文法の適用は困難である．

一方で，格文法は語順の自由を持つという特徴から，文節間の係り受け関係の有無を二値分類を用いて判断した格文法の構文木モデルを作成しても，一度モデルを生成してしまうと，その中の文節の語順が係り受け関係に基づいて限定されるため，係り受け関係の強さによって語順の制約が生まれ，格文法における語順が自由であるという特徴を生かせないといった問題も生じる．また，ある文節に着目した場合，直後の文節に対して係り受け関係を持つかどうかというアルゴリズムを用いて係り受け解析を行うため，その文節から別の文節との間に多くの文節が存在していると，それら二つの文節間に係り受け関係が存在していても，他に係り先となり得るような文節が本来，係り先の文節との間に存在することで，本来係り先となる文節以外の文節に係り受け関係が生じると判断されてしまうことで，係り受け関係の発見が困難となる．

そこで本稿では，格文法の文における係り受け関係の結合が順次可能な新しい格文法に基づく構文木モデルの提案に基づく係り受け解析手法の提案を行う．このモデルの概念は木接合文法 (TAG: Tree Adjoining Grammar) における句構造文法で表現される構文木の接合操作の概念を係り受け木の結合に利用することにより，人手によって係り受け関係のアノテーションが付与された教師データから係り受け関係の結合が行われる確率を求めることで実現できる．そのため，ある文節の係り受け関係について他の複数の文節の関係をモデルとして考慮することができ，格文法に対する新しい係り受け解析手法となり得る可能性がある．

† 同志社大学大学院， Graduate School of Doshisha University
‡ 同志社大学, Doshisha University

2. 基本的事項

本節では， Pitman-Yor 過程と 3 節で述べる TSG において構文木モデルの導出に利用する 階層 Pitman-Yor 過程について述べる．

2.1 Pitman-Yor 過程

ノンパラメトリックベイズモデルの一つである Pitman-Yor 過程 [7] は，自然言語処理の n グラム言語モデルを扱う際の利用がその一つの例として存在する．

観測された単語集合 W に含まれる単語 w が出現する n グラム分布を生成する確率過程 G を Pitman-Yor 過程 $PY(d, \theta, G_0)$ を用いて式 (1) のように表すことができる．

$$G \sim PY(d, \theta, G_0) \tag{1}$$

このとき，Pitman-Yor 過程の三つのパラメータ d, θ, G_0 はそれぞれ，d は $0 \leq d \leq 1$ の範囲を取り，観測度数を実際の度数よりも低く見積もるために用いられるディスカウント項と呼ばれるパラメータ，θ は Pitman-Yor 過程によって生成される確率過程の基底分布 $G_0 = [G_0(w)]_{w \in W}$ への依存の強さを示すパラメータ，$G_o(w)$ は観測された単語の出現頻度に基づいた n グラム分布である．

Pitman-Yor 過程はディリクレ過程にディスカウント項 d のパラメータを加えたディリクレ過程の拡張である．つまり，ディリクレ過程は無限次元のディリクレ分布として考えることができるため，Pitman-Yor 過程も ディリクレ過程と同様，無限次元のディリクレ分布を生成することのできる確率過程，すなわち加算無限性をサポートした確率過程である．Pitman-Yor 過程において $d = 0$ のとき，Pitman-Yor 過程は 式 (2) のディリクレ過程 $DP(\theta, G_0)$ と等価である．

$$G \sim DP(\theta, G_0) = \theta G_0 \tag{2}$$

式 (2) におけるディリクレ過程 G は，観測された出現単語の事象空間である r の大きさの離散空間の任意の分割に対して式 (3) で表されるディリクレ分布 $Dir(\theta G_0(w_1), \cdots, \theta G_0(w_r))$ に近似する．

$$\begin{aligned} &(G(w_1), \cdots, G(w_r)) \\ &\sim Dir(\theta G_0(w_1), \cdots, \theta G_0(w_r)) \end{aligned} \tag{3}$$

このとき，$G_0(w_r)$ は単語 w_r で基底分布 G_0 を積分した面積，すなわちその空間までの単語の出現確率の総和に相当する．

ゆえに，ディリクレ分布にディスカウント項を加えた Pitman-Yor 過程は，基底分布 G_0 に基づいて生起するあ

る単語を w_k，単語の語彙数について t としたとき，単語 w_k の出現頻度 c_k，全単語数 $c_\cdot$ を用いて式 (4) に表すことができる．

$$PY(d,\theta,G_0) = \frac{c_k - d}{\theta + c_\cdot} \cdot \frac{\theta + dt}{\theta + c_\cdot} \tag{4}$$

Pitman-Yor 過程 に対して付与されるパラメータ d,θ はノンパラメトリックであり，G_0 に近似するような分布をメトロポリス・ヘイスティングス法やギブスサンプリング等のマルコフ連鎖モンテカルロ法 [8] を用いてパラメータを収束させて推定を行う．

2.2 階層 Pitman-Yor 過程

階層 Pitman-Yor 過程 [9] は Pitman-Yor 過程を階層化した確率過程である．観測された単語の n グラム分布を基底分布とする Pitman-Yor 過程を階層化することにより，階層 Pitman-Yor 過程では深さ n において，$n-1$ の Pitman-Yor 過程を基底分布とした階層で構成される $n-1$ のフレーズの生起を条件付き確率とした単語の生起分布である n グラム分布を生成することが可能となる．加算無限性をサポートした Pitman-Yor 過程を階層化することにより， Kneser and Ney スムージング [10] を利用した n グラム分布と同様の再現性の高い言語モデルを生成することができる．

n グラム分布はある時点でのコンテキスト，すなわち特定の単語列の出現する文脈のもとで次に生起する単語が観測される確率分布をコーパスに基づいて推定することで生成される．コンテキストは Pitman-Yor 過程において添え字 u として付与され，式 (5) にこれを記す．

$$G_u \sim PY(d_{|u|},\theta_{|u|},G_{\pi(u)}) \tag{5}$$

$d_{|u|},\theta_{|u|}$ と Pitman-Yor 過程におけるパラメータに $|u|$ が付与されているのは，その時点でのコンテキストによってパラメータに最適な値を割り当てるためであり，$G_{\pi(u)}$ はその時点でのコンテキストである単語の出現分布である．

式 (5) を深さ $n-1$ のコンテキストからコンテキストが存在しない状態，すなわち基底分布に G_ϕ を獲得するまで，式 (6) に向かって再帰的コンテキストを遡る操作を繰り返すことにより， n グラム分布を生成する確率過程を生成することができる．ゆえに式 (5) は次 に収束する．

$$\sim PY(d_0,\theta_0,G_0) \tag{6}$$

G_ϕ はコンテキストが存在しない時点でのすべての単語 w と同時点での語彙数 V について $G_0(w) = \frac{1}{V}$ となるような一様分布 G_0 を基底分布として確率分布を生成するような Pitman-Yor 過程である．最終的に，パラメータ d,θ はそれぞれ一様分布と $\gamma(1,1)$ で表されるガンマ分布に収束する．これらのパラメータの合計は $2n$ を満たす．

上述のようにして生成される階層 Pitman-Yor 過程は深さ n の Suffix-Array で表現される．階層 Pitman-Yor 過程は観測された単語の出現確率によって長さ n の単語列のフレーズの生成確率を求めることができる．

3. 関連研究

3.1 二値分類による係り受け解析手法

従来の日本語の係り受け解析手法は，格文法に基づく構文木モデルの生成を文節間の係り受け関係の有無を二値分類で判定することで係り受け関係のモデルを生成し，そのモデルを解析対象の文に適用することで，解析対象文節間の係り受け関係の存在を判定している [1, 2].

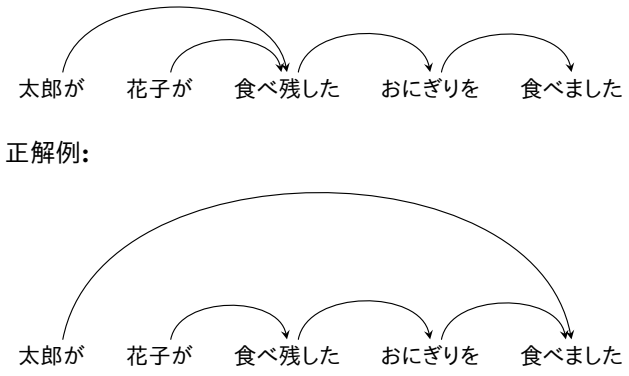

図 1: CaBoCha における構文解析の成功例と失敗例

統計的手法に基づいた日本語構文解析器の一つである CaboCha は，形態素解析器である MeCab を利用して文節を構成する語の品詞体系を抽出し，それを素性とすることで文節間の係り受け関係の有無を線形 SVM を用いて判定している [1]．つまり，対となっている文節間における係り受け関係の存在の有無は，それぞれの文節を構成する語の品詞体系に依存する．ゆえに，類似した語の品詞体系を持つ文節同士が並んでいる場合，図 1 のように係り受け関係が発生し得ない文節に対して係り受け関係が発生し，誤解析結果が出力される．

このような誤解析結果は動作の主体とその動作を表す文節の対が一文に複数存在することで，その数を増すこととなる．本研究ではこのような複雑な構造を持つ文の解析に失敗するというような問題を解決するために，木置換文法と同様，弱文脈依存性を取り入れた格文法に基づいた構文木モデルを生成し，これを利用した日本語構文解析器の開発を目指すことが，本稿における研究目的となっている．

3.2 木置換文法

TSG[11] は 文脈自由文法 (CFG: Context Free Grammar) の拡張である．一般に英語の構文木は句構造規則によって図 2 のように表されるが， TSG は基本木と呼ばれる構文木の特定のノードを，任意の深さの構文木である部分木で書き換えていくことにより，入力として与えられた文に対して対応する構文木を生成する言語モデルである．CFG が特定のノードに対して，深さ 1 の部分木である書き換え規則を利用して書き換えるのに対して， TSG は任意の深さの部分木でノードの書き換えを行っていくため，弱い文脈依存性を持つ．

TSG は $G = (T,N,S,R)$ の四つの要素によって記述される．T は非終端記号であり，N は終端記号である．また， S はルートを表すため $S \in N$ であり， R は部分木の結合規則である．TSG で用いられるモデルは基本木と部分木と呼ばれる構文木から構成され，双方の構文木のノードはともに非終端ノードと終端ノード (単語) に分けてラベル付けされている．非終端シンボルとしてラベ

ル付けされている葉ノードに対して部分木で置換していくことにより文に対する構文木の生成を行う．なお，この非終端シンボルを置換する操作を置換操作と呼ぶ．

TSG を言語モデルとして利用して構文解析を行う際は解析対象の文に対して最適な基本木を設定し，基本木の非終端シンボルに対して部分木の置換規則を用いて最適な置換操作を行うことで，文について構文木の生成を行う．S をルートノードとした基本木の非終端ノードに対し，部分木で置換を行うことで構文木を生成する．図 3 では，二つの終端記号であるそれぞれ NP に対して部分木 NP–John と NP–cookies で置換することにより構文木を生成している．

TSG の構文木モデルの獲得には階層 Pitman-Yor 過程 [12] とノードが非終端ノードが終端ノードかについてはマルコフ連鎖モンテカルロ法を組み合わせて，ランダムに部分木を分割することにより学習データに基づいたモデルの学習を行う．図 4 では構文木をランダムに部分木に分割することにより，特定の部分木のパタンを獲得している．階層 Pitman-Yor 過程とマルコフ連鎖モンテカルロ法により，学習データから基本木と部分木で構成される TSG の構文木モデルを獲得し，そうして生成されたモデルを利用することで精度の高い構文解析手法を実現している [3].

この手法は現在において英語の構文解析において最高精度を誇る手法 [6] のベースとなっている．

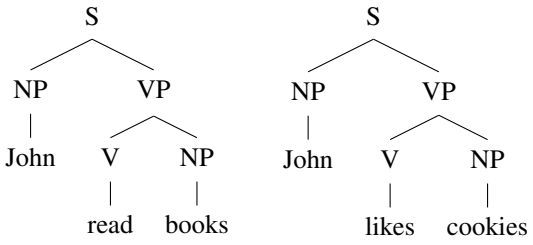

図 2: 構文木の例

3.3 木挿入文法

木接合文法 (TAG: Tree Adjoining Grammar) は計算言語学や自然言語処理でよく用いられている生成文法理論であり， Joshi[13] によって考案された形式言語である．TAG は TSG とは異なり，生成文法理論であるためそれ自体は統計的な処理を包含していない．そのため，ノードに対して行うことのできる操作を限定し，確率的モデルを付与することで TSG のような形で構文解析のための句構造規則に基づく構文木モデルを生成できるよう利用されている [3].

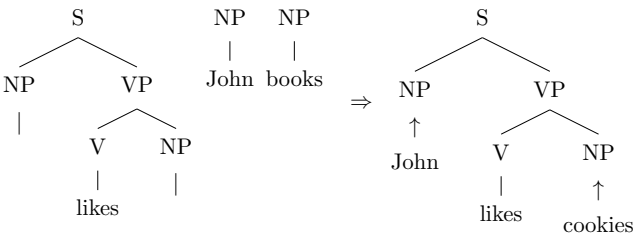

図 3: TSG による構文解析

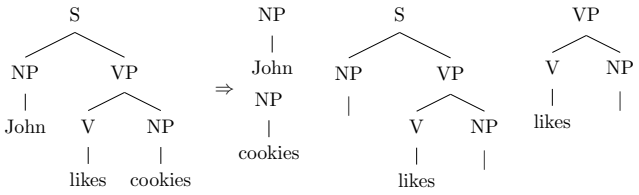

図 4: TSG に基づく構文木の獲得

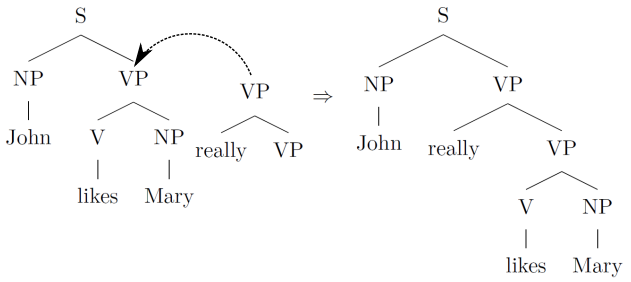

図 5: TAG による部分木挿入操作

TSG が入力された文に対して選択された基本木のノードを部分木で置換する操作を繰り返すことで構文木を生成するのに対し， TAG は図 5 のように既存の文の構文木に対して部分木の挿入操作を行うことによって，新たな構文木を生成できる．構文木のノードに対して挿入する部分木を割り当てることで，無限に長い単語長から構成される文を生成することが可能である．

ゆえに， TAG は格文法における係り受け関係の生起の特徴に近い性質であり， 4 節では TAG を拡張することによって， TSG の構文木モデルの特徴である弱文脈依存性を取り込んだ，係り受け関係の階層化とその結合を用いた格文法における構文木モデルを提案する．

4. 係り受け関係の階層化に基づいた構文木モデル

格文法を基本文法とした日本語では，文節間の係り受け関係によって文の意味役割が構成されるため，係り受け関係について以下の二つの制約のもとで，文節の出現順序が自由であるという特徴を持つ．

- 係り元の文節から複数の係り先に係り関係は発生しない
- 係り受け関係は交差しない

文節の出現順序が自由であるという特徴より，格文法は係り受け関係を構成する既存の文に対して，図 6 の文中の文節への係り受け関係の接合， 図 7 の文頭の文節への係り受け関係の接合，図 8 の文末の文節への係り受け関係の接合といった操作を行うことにより既存の文に修飾関係にある文節を図の点線で示される係り受けの係り元の文節として付与していくことで，新たな文を生成することが文法の性質上可能となっている．図 6 は「太郎は 弁当を 食べた」という文の文節「弁当を」に対して「花子の」という「弁当」の属格を表す所有格を以つ文節が係り元の文節として付与されることで「太郎は 花子の 弁当を 食べた」という文を生成し，図 7 は「太郎は 弁当を 食べた」という文の文頭の文節「太郎は」に

対して「次郎と」という主格に付与される名詞を表す文節が係り元の文節として付与されることで「次郎と 太郎は 弁当を 食べた」という文を生成している．図 8 ではそれぞれの文に，文末の文節「食べた」に対して，主格を表す「太郎は」と対格を表す「弁当を」という文節が係り元の文節として付与されることで，新な文を生成している．

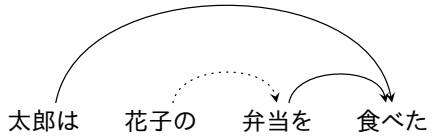

図 6: 文中の文節への係り受け関係の接合

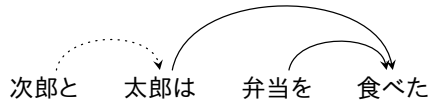

図 7: 文頭の文節への係り受け関係の接合

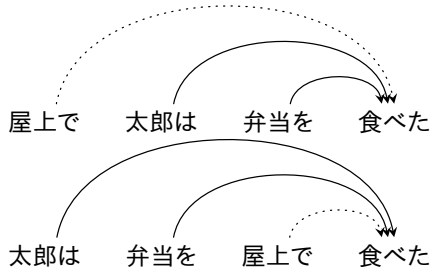

図 8: 文末の文節への係り受け関係の接合

このような格文法の特徴により，TSG を用いて，英語の構文木を生成するときと同様に句構造規則を日本語に適用すると，文節間の係り受け関係が考慮されず，格文法における語順の自由性を保証することが困難である．また，ある文から得られた部分木は他の文の構文解析に適用することができないため，日本語に対して TAG を利用しても汎用的な構文解析のモデルの生成は困難である．

そこで，本稿では係り受け構造を階層的に表現することで，その係り元の無限性について表現し，二つ以上の係り受け関係を持つような文節については後でこれらを接合することで文全体の係り受け関係を示す構文木を生成する．

上記の二つの係り受け木のうち図 6, 7 の係り受け関係を持っている文節に対する操作は，既存の係り受け関係に対する操作であるため，既存の係り受けが存在するという条件の元で生起する文節であると考えることができる．このとき，図 6, 7 の操作が行われる事象は加算無限であるといえる．よって，提案手法では末尾の文節に対して発生する係り受け関係を末尾の文節が生起する状態を起点として階層化を行うことにより，係り受け関係の階層化を取り入れたモデルを提案する．

一方で，図 8 の文末の文節への係り受け関係の接合操作は，係り先の文節に依存して係り受け関係を接合し，既存の文節の係り受け関係に対して独立である．階層化された係り受け関係は独立に生起すると考えると，これらの係り受け関係が同時に生じる確率は係り受け関係の積によって求めることができる．これにより，入力された文節の並びに対して，文節の並びから末尾の文節に基づいた係り受け関係の階層化を獲得し，獲得された係り受け関係の階層を接合することで係り受け関係の結果を出力することが可能となる．

以降，本節では係り受け関係の階層化とその接合操作の特徴を用いて，学習データから係り受け関係の階層化を利用した格文法に基づく構文木モデルを生成する手法を提案し，生成された構文木モデルを利用した新たな係り受け解析手法を提案する．

4.1 係り受け関係の階層化

従来手法のように文節間の係り受け関係の有無を二値分類で判断するのではなく，n-グラムを用いて係り受け関係の生起確率を算出し，格文法に基づいた係り受け解析のための構文木モデルを構築する．係り受け関係の生起を n-グラムの状態遷移として考えると文末の文節は，それについて係り受け関係が生じる場合，それを係り元とした係り受け関係は存在せず，必ずそれを係り先の文節とした係り元の文節が存在し状態遷移すると考えられる．そのため，文末の文節を開始状態とした n-グラムによって格文法に基づく構文木モデルを構成する．このとき， 構文木モデルを構成する n-グラムに用いる素性は文節を構成する文法要素とする．

一方で，n-グラムモデルでは n が小さいと，学習データの再現率が下がり，逆に大きいと状態数が爆発的に増加し，モデルのサイズが大きくなってしまう．n グラムを用いた構文木モデル生成の際に，$n = 2$ とすると二つの文節間の文法要素を素性として文節間に係り受け関係が生じる確率をモデルに取り入れることになり，係り受け関係を二つの文節間のみしか考慮しているため， 3.1 節で挙げた問題点に関して 2 -グラムでは従来手法との変わりがない．

そこで，本手法では階層 Pitman-Yor 過程を拡張することで確率過程の階層の深さ n を可変長で扱える可変長階層 Pitman-Yor 過程 [14] を利用する．これより，任意の長さの文節によって構成される係り受け関係の生じる確率をモデルとして扱うことができる．文末の文節を根として任意の深さで，文末の文節に係り受け関係を持つ文節を係り元の文節とし，次にその文節が生じる確率を基底分布として係り元の文節が生じる確率を求める場面で，可変長階層 Pitman-Yor 過程を利用することにより，係り受け関係を階層化したモデルを生成することができる．また，文末の文節に基づいて階層化した係り受け関係は係り元の文節が生じない，つまり，係り受け関係がその文節で終了する確率について割り当てることで，文末の文節に基づいて係り受け関係の解析を可能にする．

図 9 では係り受け木で表現されている「次郎と 太郎は 花子の 弁当を 屋上で 食べた」という文から，文末の文節「食べた」に基づいて係り受け関係の階層化を行うことで，任意の長さの三つの係り受け木をモデルとして取得している．このとき， <s> はそれ以上係り受け関係が生じない場合，即ち文頭を表し， </s>は 文末を表す記号である．文末の文節に基づき階層化された係り受け関係は文末の文節が係り受け関係に複数の係り元の文節を持つ場合， 4.2 節で解説するこれらが同時に生起する

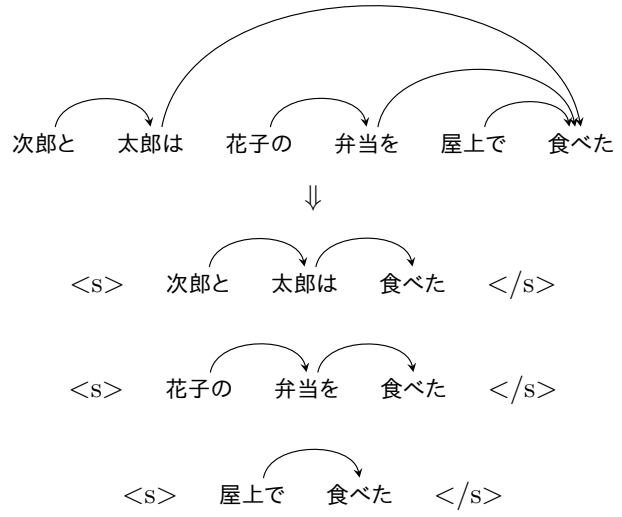

図 9: 係り受け関係の階層化

確率について考えることで，文末に対する係り受け関係を利用する．

4.2 階層化した係り受け関係の接合による係り受け解析

4.1 節で示した係り受け関係の階層化により，任意の長さの文節から構成される係り受け関係のモデルを生成する．これにより，入力された文に対して，文末の文節を係り先とする係り受け関係を取得することができる．

階層化した係り受け関係から構築されたモデルは図 10 にあるように図 9 の過程で生成したモデルを接合することで,「太郎が 花子の 弁当を 食べた」というような係り受け木を出力し，この文の係り受け解析を行うことができる．このとき，他にも係り受け関係の結合パタンは存

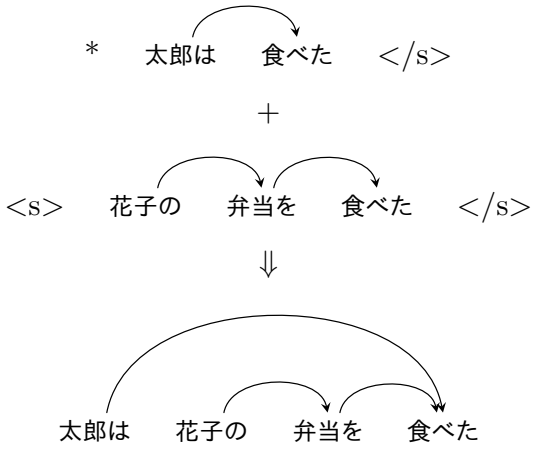

図 10: 階層化された係り受け木の統合

在するが，係り受け関係を結合する確率についてはそれぞれの係り受け関係の生起確率の積によって求めることにより，尤もらしい係り受け解析結果を出力する．これにより，既存の係り受け解析手法における問題点である語順の自由が制約されるために，係り受け解析の誤解析が生じる問題の解決を行うことで，誤解析を防ぐものである．

5. 評価実験

1995 年度の毎日新聞のデータ*に対して様々な言語情報が人手で付与された京都大学テキストコーパス†は形態素，文節間の係り受け関係等が示されている日本語コーパスの一つであり，形態素解析や係り受け解析といった自然言語処理の基礎的なタスクに利用される．

日本語における一般的な係り受け解析器 CaboCha は京都大学テキストコーパスに含まれる文節間の係り受け関係のアノテーションが付与された文を学習データとして利用することにより，係り受け木のモデルを生成している係り受け解析器の一つである [1]．本節では提案手法の係り受け関係の階層化に基づいた構文木モデル生成の際に， CaboCha と同様に京都大学テキストコーパスを学習データとして利用し，生成したモデルによる係り受け解析の精度について CaboCha と比較することにより評価を行った．

5.1 係り受け関係の階層化に基づいた構文木モデルの実装

本節では提案手法である係り受け関係の階層化に基づいた構文木モデルの生成とその利用について記述する．学習データとしては上述した通り京都大学テキストコーパスを利用するが，評価実験のモデル生成の際に使用した学習データは 1995 年 1 月 1 日分のデータを利用した．モデルの生成では，各文の係り受け関係を文末の文節に基づいた各文節の品詞体系からなる係り受け関係の構造を抽出し，各係り受け関係にういて文末の文節を根として階層化し構文木モデルを生成した．このとき，可変長階層 Pitman-Yor 過程のパラメータ推定には Gibbs イテレーションを 100 回繰り返し，モデルを生成した．

生成した構文木モデルを利用した構文解析手法は第一に入力された文を形態素解析した後に文節単位に分割する．次に文節を構成する形態素の品詞体系で構成される文節の並びを構文木の係り受け関係を階層化したモデルに含まれる可変長の係り受けと照合する．これにより，入力された文の文末の文節を根として生起する係り受け関係を抽出した後に接合可能な係り受け関係について各文節を根とした係り受け関係について発見していくことで，係り受け解析を行う．

5.2 結果

テストデータは学習データの 1995 年 1 月 1 日のデータから 1 割を抽出し，文節間の係り受け関係は一般的に係り受け解析手法の評価に利用される式 (7)[1] を利用して係り受け解析の精度を測定した．テストデータに対して，文節間の係り受け関係について二値分類を用いて係り受け関係の有無を判断することで文全体の係り受け解析を行う CaboCha を用いた既存手法と，提案手法によって作成した構文木モデルを利用した係り受け解析手法を利用する．

*CD-毎日新聞データ集 http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

†京都大学テキストコーパス http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス

$$
\begin{aligned}
X &= \text{各手法によって得られた文節間の係り受け関係と正解データの係り受け関係の一致数} \\
Y &= \text{テストデータの文節間の係り受け総数} \\
\text{正解率} &= \frac{X}{Y} \qquad (7)
\end{aligned}
$$

表 1 の実験結果より，提案手法が従来手法と比較して精度が劣るということが分かった．これは従来手法が文節の格を用いているのに対し，提案手法は文節の品詞体系のみを用いてモデルを構築し，これを利用して係り受け解析を行っているためであるといえる．したがって従来手法と提案手法との間に精度の差が開いたものであると考えられる．

	従来手法	提案手法
正解率 (%)	87.1	76.9

表 1: 実験結果

6. おわりに

本稿では日本語における係り受け解析手法において係り受け関係を階層化し，構文木の更新を考慮した係り受け木のマージを考えることで，係り受け関係の発見に弱文脈依存性を取り入れ，既存の係り受け解析手法において，語順と語の距離が考慮されていない問題の解決を図った．

今後の課題としては，評価実験の結果を踏まえ，係り受け関係の階層化を利用した構文木モデルの生成の素性として，学習データに含まれている格構造を取り入れ生成するモデルの正確性を向上させ，係り受け解析精度の改善を目指す．また，係り受け関係がアノテーションされた京都大学テキストコーパスの全データを用いた評価実験を行う必要がある．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金挑戦的萌芽研究 (課題番号: 25540150) の支援による．ここに記して謝意を表す．

参考文献

[1] 工藤拓, 松本裕治. チャンキングの段階適用による係り受け解析. 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2001, No. 20, pp. 97–104, 2001.

[2] 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理, Vol. 14, No. 4, pp. 67–81, 2007.

[3] Trevor Cohn, Phil Blunsom, and Sharon Goldwater. Inducing tree-substitution grammars. *Journal of Machine Learning Research*, Vol. 11, pp. 3053–3096, 2010.

[4] Matt Post and Daniel Gildea. Weight Pushing and Binarization for Fixed-Grammar Parsing. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pp. 89–98. Association for Computational Linguistics, 2009.

[5] Phil Blunsom and Trevor Cohn. Unsupervised induction of tree substitution grammars for dependency parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (EMNLP '10), pp. 1204–1213. Association for Computational Linguistics, 2010.

[6] Hiroyuki Shindo, Yusuke Miyao, Akinori Fujino, and Masaaki Nagata. Bayesian symbol-refined tree substitution grammars for syntactic parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pp. 440–448. Association for Computational Linguistics, 2012.

[7] Jim Pitman and Marc Yor. The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *The Annals of Probability*, Vol. 25, No. 2, pp. 855–900, 1997.

[8] 北川源四郎, 有川節夫, 小西貞則, 宮野悟編. 計算統計学の方法: ブートストラップ・EM アルゴリズム・MCMC. シリーズ予測と発見の科学. 朝倉書店, 2008.

[9] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pp. 985–992. Association for Computational Linguistics, 2006.

[10] R. Kneser and H. Ney. Improved backing-off for M-gram language modeling. In *Acoustics, Speech and Signal Processing*, Vol. 1, pp. 181–184, 1995.

[11] Trevor Cohn and Mirella Lapata. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research (JAIR)*, Vol. 34, pp. 637–674, 2009.

[12] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pp. 985–992. Association for Computational Linguistics, 2006.

[13] Aravind K. Joshi, Leon S. Levy, and Masako Takahashi. Tree Adjunct Grammars. *Journal of Computer and System Sciences*, Vol. 10, No. 1, pp. 136–163, 1975.

[14] 持橋大地, 隅田英一郎. Pitman-yor 過程に基づく可変長 n-gram 言語モデル (言語モデル・応用). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2007, No. 35, pp. 63–70, mar 2007.