情報処理学会論文誌　データベース　Vol.7　No.2　61–69 (June 2014)

# 部分統語構造を考慮した階層的確率オートマトンに基づく教師なしチャンキング

若林 啓[1,a)]

受付日 2013年12月22日, 採録日 2014年3月31日

**概要**：チャンキングは，単語の系列から名詞句や前置詞句といった浅い統語構造を抽出する技術であり，固有表現抽出や機械翻訳などで重要な前処理であると考えられている．これまでに提案されている多くのチャンキング手法は教師あり学習に基づいており，教師データに現れない文章表現を多く含む Web 上の文書には適用が難しい．本研究では，依存構造解析モデルの平坦近似に基づいた平坦近似依存文法モデル（FADG）を用いることで，チャンク間の局所的な統語構造を考慮した教師なしチャンキング手法を提案する．FADG は線形鎖モデルを階層的に接続した階層型隠れマルコフモデル（HHMM）の枠組みで形式化するため，HHMM の効率的な教師なし学習アルゴリズムを適用できる．実験により，提案モデルが局所的な統語構造を効果的に推定し，これによって高い精度で教師なしチャンキングを行えることを示す．

**キーワード**：教師なしチャンキング，部分構文解析，階層型隠れマルコフモデル

# Unsupervised Chunking with Partial Syntactic Analysis Using Hierarchical Probabilistic Automaton

Kei Wakabayashi[1,a)]

Received: December 22, 2013, Accepted: March 31, 2014

***Abstract:*** Chunking is a natural language processing task to extract shallow syntactic structures like noun phrases or prepositional phrase, and it plays an important role in various applications such as named entity extraction and machine translation. Most chunking algorithms proposed so far are based on supervised learning, but they depend on the domain of supervision documents that often consist of news articles and are not effective for analyzing Web documents or microblogs. In this paper, we propose an unsupervised chunking method based on the *Flat Approximated Dependency Grammar model* (FADG) to capture local syntactic dependency structures between chunks. The FADG is formalized as a Hierarchical Hidden Markov Model (HHMM) and we can conduct the unsupervised learning of FADG efficiently by using a sophisticated inference algorithm for HHMMs. The experimental results show the effectiveness of the proposed method in chunking accuracy comparison.

***Keywords:*** unsupervised chunking, partial parsing, Hierarchical Hidden Markov Model

## 1. はじめに

近年，インターネットや企業内データベースなどにおいて大量の電子化された文書が蓄積されるにともない，自然言語データの解析技術が重要性を増してきている．特に，固有表現抽出や関係抽出，キーワード抽出など，文書集合から構造化された情報を抽出する処理は，巨大な文書集合からより価値の高い知識を獲得するためのコア技術といえる．これらの処理では自然言語文章の統語解析が重要な役割を果たすが，とりわけ単語系列から実体（entity）などを表す意味単位を抽出するチャンキング処理を高い精度で行うことが必要不可欠である．

チャンキングは，統語的に最も密接につながった部分単語系列を抽出する技術である．構文解析において最も単語

[1] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[a)] kwakaba@slis.tsukuba.ac.jp

に近い部分構文木の推定に対応することから，チャンキングは浅い構文解析（shallow parsing）とも呼ばれる．機械翻訳や固有表現抽出では，浅いレベルの統語関係の同定が結果に大きな影響を与える．たとえば，“New York stock exchange” は 1 つの実体を表しており，それぞれの単語を独立に扱うべきではない．完全な構文木の推定を 100%の精度で行うことは難しいため，このような応用では特にチャンクの同定精度に焦点を当てた統語解析手法が求められる．また，多くの完全構文解析手法は系列長の 3 乗のオーダの計算量が必要であり，実行時間の観点でもチャンキングに特化した手法を考えることの意義がある．

これまでに提案されているチャンキング手法は，教師あり手法と教師なし手法に大きく分けられる．教師あり手法では，人手で作成したチャンクの正解データを利用して，隠れマルコフモデル（HMM）[4] や条件付き確率場 [9]，Support Vector Machine [11] といった分類器や系列ラベリングモデルの教師あり学習を行い，これに基づいてチャンキングを行う．教師あり学習手法は，教師データと同じドメインの文章では高い精度でチャンクを同定できる反面，教師データのドメインに依存するという問題がある．正解チャンクが付与されたコーパスには新聞記事を用いたものが多いが，教師データには現れないタイプの固有表現，新聞記事では使わない言い回し，くだけた文章，あるいは異なる言語などが含まれた想定外の入力文章が与えられると，教師あり手法は極端に精度が悪化する．このことは，Web 文書や Twitter などの解析では特に深刻な問題になる．

一方，教師なし手法は，チャンクの正解データが付与されていない言語データのみを用いてモデルを構築し，チャンキングを行う．人手によるアノテーションが必要ないため，チャンキング対象の文章も含め，場合によっては無尽蔵に学習データを利用できるが，教師なし学習によってチャンクの概念に対応した知識を獲得できるようなモデルが必要になる．Pate ら [6] は，正解チャンクが利用できない代わりに，音韻情報を利用してチャンクの手がかりを得るアプローチを提案している．ここでは，音韻情報が付与された学習データを用いて HMM の拡張モデルを学習し，得られたパラメータを用いて文章のチャンキングを行う．音韻情報は有効な手がかりと考えられるが，音声情報が利用できるドメインは限られる．ドメインに依存しないチャンキングの実現のためには，テキストデータのみから学習できる手法が望ましい．

Ponvert ら [7] は，単語系列のみの学習データを用いて，高い精度でチャンキングを行う手法を提案している．本論文では，文献 [7] の提案モデルについて考察し，局所的な統語構造の考慮による精度向上の可能性について議論する．この考察に基づいて，教師なし構文解析で用いられる依存構造解析モデルを局所的な統語構造解析のために平坦近似し，階層型隠れマルコフモデルの枠組みによってチャンキングモデルとシームレスに組み合わせる手法を提案する．

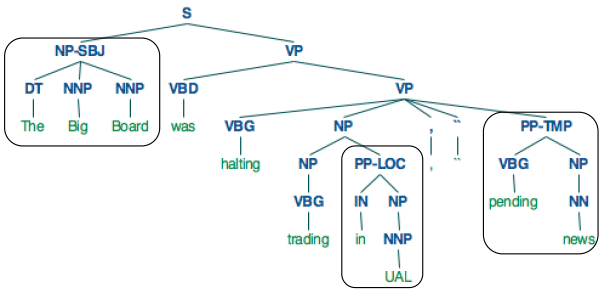


図 1　構文木に基づくチャンクの定義
**Fig. 1**　Chunk definition based on parse tree.

本論文の構成は以下のとおりである．2 章では，チャンキング問題の定義を述べる．3 章で，文献 [7] を含む関連研究について述べる．4 章で，提案する部分統語構造を考慮したチャンキング手法について述べる．5 章で実験結果を示し，6 章で結論を述べる．

## 2.　チャンキング

チャンキングは，与えられた系列長 T の単語系列 $w_1, \ldots, w_T$ から，ある目的に従って，重複のない区間集合 $\{c_1, \ldots, c_N | c_i = (t_i, s_i), 1 \leq t_1 < s_{i-1} < t_i < s_i < t_{i+1} < s_N \leq T\}$ を抽出する処理であり，区間 $c_i$ をチャンクと呼ぶ．正解チャンクの定義は文献 [7] に従い，与えられた正しい構文木において，以下の条件をすべて満たすノードが対応する系列区間をチャンクと定義する．

(1) 2 語以上の単語を葉ノードに持つ．
(2) 子ノードに条件 (1) を満たすノードが含まれていない．

本定義によるチャンクの例を**図 1** に示す．“in UAL” や “pending news” のチャンクは NP（名詞句）を子ノードに含んでいるが，それぞれ “UAL”，“news” の 1 語だけであるため条件 (2) に抵触せずチャンクとなる．もしこの NP が “the news” のように 2 語であれば，“the news” がチャンクになる．

教師ありチャンキングでは品詞タグが付与された単語系列を入力とする場合もあるが [4]，本研究では，単語の系列のみに基づいてチャンクを推定する問題を扱う．同様に，チャンクの品詞種類も出力とする場合もあるが，本論文ではチャンク区間のみを出力とする．

## 3.　関連研究

### 3.1　HMM による教師なしチャンキング

チャンキング問題は，単語系列に対してチャンク区間を表現する IOB タグを付与する系列ラベリング問題に帰着する．IOB タグ形式では，それぞれの単語について，チャンクの開始を表す B タグ，チャンクの継続を表す I タグ，どのチャンクにも含まれていないことを表す O タグのいず

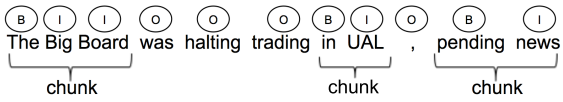


図 2　IOB タグによるチャンク区間表現
**Fig. 2**　IOB tag expression of chunk segment.

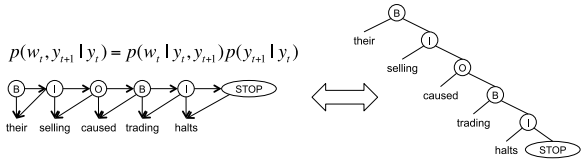


図 3　IOB-PRLG モデル
**Fig. 3**　IOB-PRLG model.

れかを付与する．IOB タグを付与することで，チャンク集合は唯一に決まる（**図 2**）．このため，教師なしチャンキングは，系列ラベリングモデルの教師なし学習を用いることによって実現できる．

隠れマルコフモデル（HMM）は，単語系列 $w_1, \ldots, w_T$ の単語に 1 対 1 で対応した状態の系列 $y_1, \ldots, y_T$ を潜在変数として持つ生成モデルである．単語は当該時刻の状態にのみ依存した確率分布に従い（$p(w_t|w_{1:t-1}, w_{t+1:T}, y_{1:T}) = p(w_t|y_t)$），状態系列には単純マルコフ性が仮定される（$p(y_t|y_{1:t-1}) = p(y_t|y_{t-1})$）．ここでは，状態の集合を $\{I, O, B\}$ として定義する．チャンクは 2 単語以上から構成されるため，$y_t = B$ のとき $y_{t+1} = I$ である．また，チャンクは B タグから開始するため，$y_t = O$ のとき $y_{t+1} \neq I$ である．これらの制約は，状態遷移確率について $p(y_{t+1} = I|y_t = B) = 1$，$p(y_{t+1} = I|y_t = O) = 0$ の制約を与えることに対応する．ここでは，この制約を加えた HMM を IOB-HMM と呼ぶ．

HMM は生成モデルであるため，単語系列の集合と乱数で初期化されたパラメータ集合を与えることで EM アルゴリズムによる教師なし学習を行うことができる．HMM の EM アルゴリズムとして，単語系列長 $T$ について線形の計算量 $O(T)$ で実行できる Baum-Welch アルゴリズムが知られている [8]．また，学習されたパラメータを与えることで，最尤の状態系列を $O(T)$ で求める Viterbi アルゴリズムがある．教師なしチャンキングは，チャンキング対象の単語系列集合 $D$ を用いて Baum-Welch アルゴリズムを実行し，学習されたパラメータを用いて $D$ について Viterbi アルゴリズムを実行することで実現できる．

### 3.2　IOB-PRLG 手法

Ponvert ら [7] は，HMM を拡張した確率モデルを用いた教師なしチャンキング手法を提案している．この拡張モデルでは，単語 $w_t$ は，状態 $y_t$ と $y_{t+1}$ に依存した確率分布 $p(w_t|y_t, y_{t+1})$ に従って生成される．このモデル化は，確率文脈自由文法（Probabilistic Context Free Grammar，PCFG）[12] の特殊な場合と等価である．PCFG の文法がチョムスキー標準形であり，かつ導出規則右辺の第 1 項が単語である規則 $y_t \rightarrow w_t y_{t+1}$ のみを含む場合，**図 3** に示すように右側に偏った構文木が生成される．このように制約された PCFG は，Probabilistic Right Linear Grammar（PRLG）と呼ばれる．Ponvert ら [7] の拡張モデルは PRLG モデルと等価な表現能力を持つことから，本論文ではこの手法を IOB-PRLG と呼ぶ．

PRLG の潜在変数は線形鎖であるため，Baum-Welch アルゴリズムの単純な拡張によって計算量 $O(T)$ で推論および教師なし学習を行うことができる．文献 [7] では，教師なし学習によって得られたパラメータを用いて状態系列の Viterbi 推定を行うことで，高い精度で IOB タグを推定できることを示している（表 1）．

IOB-PRLG では，次の IOB タグを考慮して単語の生成確率が決まることから，チャンクの開始に対応する単語の確率分布 $p(w_t|y_t = B, y_{t+1} = I)$，チャンクの中間に対応する確率分布 $p(w_t|y_t = I, y_{t+1} = I)$，終わりに対応する確率分布 $p(w_t|y_t = I, y_{t+1} = O)$ を，それぞれ独立したパラメータとして学習することができる．このモデル化が HMM の結果を大きく上回っていることから，チャンクの開始，中間，終了にそれぞれ特徴的な単語の分布が存在するという仮定が有効であると考えられる．このことは Pate ら [6] でも指摘されており，ここではチャンクの終わりに対応する E タグを IOB タグセットに追加した OBIE タグセットを提案している．本研究ではこのアイデアに従い，B，I，E のタグを用いてチャンクの確率分布を考える．

IOB-PRLG は精度の高い教師なしチャンキングモデルであるが，抽出した部分系列どうしの統語関係を考慮できないことによる推定誤りを起こす．たとえば，図 1 の文章に対して，IOB-PRLG は “halting trading” の部分系列をチャンクとして抽出する．“halting trading” はある程度の頻度で現れる特徴的な部分系列ではあるが，この 2 単語は続く単語列を含めて動詞句を構成していることから，定義のうえでは条件 (2) に反しておりチャンクではない．同様の誤ったチャンクが動詞句（“did n’t dump” など）や to 不定詞（“to sell” など）といった，特徴的な部分系列について多く抽出される．これらの抽出も場合によっては有用ではあるが，チャンキングを他の自然言語処理応用のモジュールとして考えたとき，定義上のチャンクとしての正確さを基準に議論することは重要である．

これらの擬陽性を取り除くには，チャンク候補どうしの統語関係を考慮する必要がある．本研究では，チャンク候補となる部分系列の局所的な統語関係を考慮することで，チャンクの定義を満たさない候補を除去するアプローチを考える．

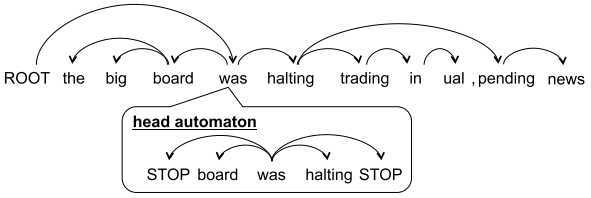


図 4　DMV による依存構造解析
**Fig. 4**　Dependency parsing by DMV.

### 3.3　教師なし構文解析

チャンキングが浅い統語構造を抽出するのに対して，教師なし構文解析では，完全な構文解析木を教師データを用いずに推定することを目的とする．教師なし構文解析の代表的なモデルとして，Dependency Model with Valence（DMV）が知られている [3]．DMV は依存構造解析（dependency parsing）に基づく確率モデルであり，主辞（head）に依存して項（argument）が生成される過程をモデル化する（**図 4**）．各項は必ず 1 つの主辞に依存して生成されるが，各主辞は 0 個以上の複数の項を生成する．主辞 $h$ は，矢印の方向 $d \in \{left, right\}$ それぞれについて，項を生成するかしないかを選択する確率分布を持つ．項を生成しないことの選択は擬似的に STOP シンボルの生成として表現され，STOP シンボルが生成されるまで項が生成される．主辞に依存して各方向に項の系列を生成する部分確率モデルは主辞オートマトン（head automaton）と呼ばれる．

項として生成された単語は再帰的に主辞となり，当該の単語に依存した主辞オートマトンに従って項を生成する．DMV は等価な確率文脈自由文法（Probabilistic Context Free Grammar，PCFG）に変換することができるため，Inside-Outside アルゴリズム [12] を用いて推論および教師なし学習を行うことができる．Inside-Outside アルゴリズムの計算量は，系列長 T に対して $O(T^3)$ であるため，系列長が長いと計算時間が非常に大きくなる．

教師なし構文解析の解析結果からチャンク部分を抽出することによって教師なしチャンキングを行うこともできるが，計算量の問題により大規模文書に適用することは現実的に難しい（表 4）．チャンキング問題に特化した手法を用いることにより，計算量を系列長に対して線形のオーダに抑え，大規模なテキストデータに対して現実的な実行時間で適用することができる．この点において，教師なし構文解析とは別に教師なしチャンキング手法を考えることの有用性がある．

## 4.　平坦近似依存文法による教師なしチャンキング

本研究では，依存構造解析モデルを平坦近似することにより，局所的な統語構造に基づいてチャンクを抽出する手法を提案する．チャンクの定義から，構文木において他のチャンクを入れ子状に含んでいるノードに対応する部分系列は，チャンクではない．これは言い換えれば，チャンクに対応する部分系列に含まれる単語は，主辞としてチャンク外の他の単語を生成しないことを意味する．

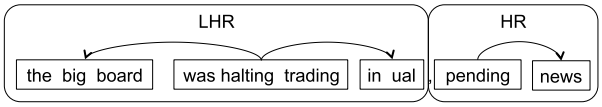


図 5　局所的な依存構造に基づく主辞オートマトン系列表現
**Fig. 5**　Head automaton sequence for local dependency structure.

このため，IOB タグを主辞と項で細分化することで，チャンク抽出の精度向上が期待できる．しかし，主辞と項の前後関係の特性を単純マルコフ連鎖で表現することは難しい．本研究では，従来手法で用いられてきた線形鎖モデル構造を拡張し，もう 1 段階上位の統語構造を考慮した階層型隠れマルコフモデル（Hierarchical Hidden Markov Model，HHMM）[5] を用いることで，チャンクの推定精度の向上を目指す．本章では，局所的な依存構造解析が HMM の階層的な接続で近似的に表現可能であることを示し，主辞と項のいずれかの値をとる潜在変数系列を上位階層に対応させた HHMM（図 8）を構築する方法について述べる．また，このモデルの推論に基づいたチャンク抽出方法について述べる．

### 4.1　依存構造解析モデルの平坦近似

ここでは，単語単位の依存構造ではなく，部分系列を単位とした依存構造を考える（**図 5**）．部分系列単位の分割が仮に与えられているとし，以下ではこの部分系列単位を**ブロック**と呼び，チャンクの候補とする（ブロックに対応する部分系列の分割が与えられているという仮定は最終的には取り除かれる）．本手法では，系列中で隣り合うブロック間に限定した，局所的な主辞と項の関係を推定する．依存構造に対して以下の 2 つの制約を考える．

(1) 局所性．主辞オートマトンが生成できる項は，系列上で隣接したブロックのみである．
(2) 非再帰性．項として生成されたブロックは，再び主辞として別の項を生成しない．

制約 (1) により，主辞オートマトンは (i) 左と右に項を持つ場合（LHR），(ii) 左にのみ項を持つ場合（LH），(iii) 右にのみ項を持つ場合（HR）の 3 通りになる．また，完全な依存構造において隣接していないブロックを主辞として持つ項は，依存関係が分断された孤立した項（Arg）となる．

制約 (2) により，ブロックは主辞か項のどちらかの役割を持つ．このため，ブロックの役割を潜在変数としてモデル化し，モデル推論によって主辞であるブロックをチャンクから除外することを目指す．また，再帰性を持たないことにより，文章全体を主辞オートマトンの系列と見なせ

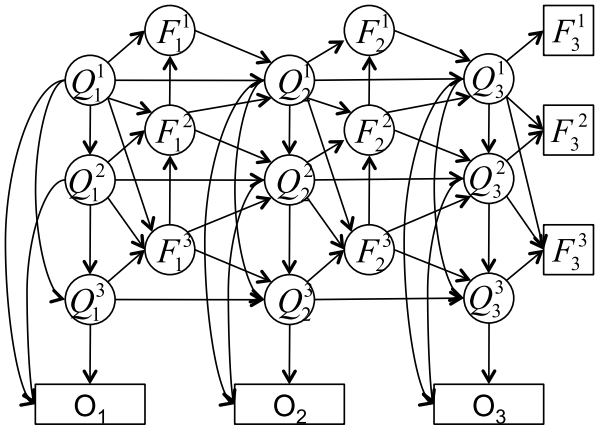


図 6　階層型隠れマルコフモデルのグラフィカルモデル
**Fig. 6**　Graphical model for hierarchical hidden Markov models.

る．図 5 に，制約を満たすような主辞オートマトンの系列表現を示す．この図では，“was halting trading” を主辞とした局所依存構造 LHR と，“pending” を主辞とした構造 HR が系列を構成している．提案手法では，この構造ラベル自体を潜在変数とした確率オートマトンを構築する．これは，本来再帰的に深い階層を持つ依存構造から一次元の系列構造への一種の射影である．この射影に対応した依存構造解析モデルから確率オートマトンへの変換を，依存構造解析モデルの**平坦近似**と呼ぶ．

### 4.2　平坦近似依存構文モデル

ここでは，平坦近似した依存構造を上位階層に対応づけた階層型隠れマルコフモデル（HHMM）を定義する．**図 6** に，HHMM のグラフィカルモデルを示す．潜在変数 $Q_t^d$ は，時刻 $t$，階層 $d$ での隠れ状態である．階層 $d=1$ は最上位階層を，$d=D$ は最下位階層を意味する．また，潜在変数 $F_t^d$ は，マルコフ連鎖の終了を表す 2 値の確率変数である．$F_t^d=1$ のとき，階層 $d$ のマルコフ連鎖が時刻 $t$ で終了したことを意味する．階層 $d$ の状態は，階層 $d+1$ のマルコフ連鎖が終了しなければ遷移できない（$Q_t^d = Q_{t-1}^d$ if $F_{t-1}^{d+1}=0$）．終了したマルコフ連鎖は，次の時刻で再び上の階層の状態に依存して初期化される．HHMM の推論とは，$Q_t^d$ と $F_t^d$ の値を推定することである．$F_t^d$ の推定はマルコフ連鎖の分断を推定することに対応しているため，HHMM の推論は一種の系列セグメンテーションとして働く．このため，依存構造の部分系列単位であるブロックに対応する部分系列の分割は，モデル推論によってシームレスに推定される．

HHMM はパラメータとして初期状態確率分布 $\pi_k^d$，遷移確率分布 $A_k^d$，出力確率分布 $B_k$ を持つ．これらのパラメータは以下の確率分布に対応する．

$$\pi_k^d(i) = p(Q_t^d = i | F_{t-1}^d = 1, Q_t^{1:d-1} = k)$$
$$A_k^d(i,j) = p(Q_t^d = j | Q_{t-1}^d = i, F_{t-1}^{d+1} = 1, F_{t-1}^d = 0, Q_t^{1:d-1} = k)$$
$$B_k(v) = p(w_t = v | Q_t^{1:D})$$

ここでは，状態の集合を以下のように対応づけて定義する．最上位階層 $d=1$ では，3 種類の主辞オートマトン LHR，LH，HR および依存関係の分断された項 Arg と，O タグ，カンマやピリオドなどの文章の区切り記号のみを生成する状態 Stop の 6 種類の状態を定義する．

$$Q^1 \in \{LHR, LH, HR, Arg, O, Stop\}$$

階層 $d=2$ には，階層 1 に対応した主辞オートマトンの状態が対応する．

$$Q^2 \in \{L, H, R, Arg, O, Stop\}$$

階層 $d=3$ には，階層 2 の状態に応じたブロックの確率分布を表現する HMM が対応する．ここでは，Pate ら [6] の OBIE タグに基づいて，ブロックの開始（B），中間（I），終了（E）に対応するの 3 種類の状態と，1 単語からなるブロック（S）に対応した状態を持つ HMM を考える．以降では，この HMM を BIES オートマトンと呼ぶ．

ここでは，階層 2 の状態が主辞の場合と項の場合で，2 種類の独立したパラメータを持つ BIES オートマトンによってブロックの生成確率を与える．このモデル化は，主辞オートマトンにおいて，項の確率分布が主辞の単語に依存しないという近似に対応している．このモデル化によって，主辞になりやすいブロックの確率分布と，項になりやすいブロックの確率分布を独立のパラメータとして学習する．このため，階層 $d=3$ の状態集合には，ブロックが主辞である場合の BIES オートマトンの状態 $B_H$，$I_H$，$E_H$，$S_H$ と，項である場合の BIES オートマトンの状態 $B_A$，$I_A$，$E_A$，$S_A$ が含まれる．加えて，O タグと Stop タグに対応した状態も含む．

$$Q^3 \in \{B_H, I_H, E_H, S_H, B_A, I_A, E_A, S_A, O, Stop\}$$

これまでの議論に基づいた確率モデルを表現するため，モデルパラメータを制約する．記法 $\pi_k^d(i=X)$ は，$i$ に $X$ 以外の値をとるとき，記法 $\pi_k^d(i \in \{X_1,\ldots,X_n\})$ は $i$ に $\{X_1,\ldots,X_n\}$ 以外の値をとるとき，それぞれ $\pi_k^d(i)=0$ という制約を意味する．また，記法 $A_k^d(i=X, j=Y)$ は，$j$ に $Y$ 以外の値をとるとき，記法 $A_k^d(i=X, j\in\{Y_1,\ldots,Y_n\})$ は $j$ に $\{Y_1,\ldots,Y_n\}$ 以外の値をとるとき，それぞれ $A_k^d(i=X,j)=0$ という制約を意味する．

階層 2 のマルコフ連鎖は，階層 1 に従って決定的に決まる．たとえば，階層 1 の状態が LH である場合，階層 2 の初期状態は確率 1 で状態 L であり，状態 L からの状態遷移は確率 1 で状態 H，また状態 H からの遷移は確率 1 でマルコフ連鎖の終了を表す擬似状態 End である．上記の記法を用いると，パラメータ制約は以下のように定義できる．

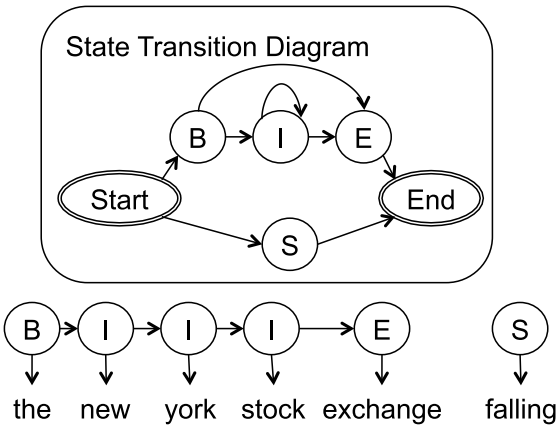


図 7　BIES オートマトン
**Fig. 7**　BIES automaton.

$\pi_k^2(i = L), A_k^2(i = L, j = H), A_k^2(i = H, j = R),$
$A_k^2(i = R, j = End)$ `if` $Q^1 = LHR$

$\pi_k^2(i = L), A_k^2(i = L, j = H), A_k^2(i = H, j = End)$
`if` $Q^1 = LH$

$\pi_k^2(i = H), A_k^2(i = H, j = R), A_k^2(i = R, j = End)$
`if` $Q^1 = HR$

$\pi_k^2(i = Arg), A_k^2(i = Arg, j = End)$ `if` $Q^1 = Arg$

$\pi_k^2(i = O), A_k^2(i = O, j = End)$ `if` $Q^1 = O$

$\pi_k^2(i = Stop), A_k^2(i = Stop, j = End)$ `if` $Q^1 = Stop$

階層 3 のマルコフ連鎖は，BIES オートマトンの特性を表現するため，**図 7** に示す状態遷移図において矢印のない状態遷移の確率が 0 になるようにモデルパラメータを制約する．制約は以下のように定義できる．

$\pi_k^3(i \in \{B_H, S_H\}), A_k^3(i = B_H, j \in \{I_H, E_H\}),$
$A_k^3(i = I_H, j \in \{I_H, E_H\}), A_k^3(i = E_H, j = End),$
$A_k^3(i = S_H, j = End)$ `if` $Q^2 = H$

$\pi_k^3(i \in \{B_A, S_A\}), A_k^3(i = B_A, j \in \{I_A, E_A\}),$
$A_k^3(i = I_A, j \in \{I_A, E_A\}), A_k^3(i = E_A, j = End),$
$A_k^3(i = S_A, j = End)$ `if` $Q^2 \in \{L, R, Arg\}$

$\pi_k^3(i = O), A_k^3(i = O, j = End)$ `if` $Q^2 = O$

$\pi_k^3(i = Stop), A_k^3(i = Stop, j = End)$ `if` $Q^2 = Stop$

単語の生成確率分布は，最下層の状態 $Q^3$ にのみ依存して決まる．ただし，$Q^3 \neq Stop$ の場合は，以下の punctuation set に含まれる単語の生成確率は 0 に，$Q^3 = Stop$ の場合は，punctuation set に含まれない単語の生成確率は 0 に制約する．

．　？　！　；　，　－－

この制約によって，punctuation によって必ずチャンクが区切られる制約を実現する．punctuation set の定義は，Ponvert ら [7] の記述に従っている．

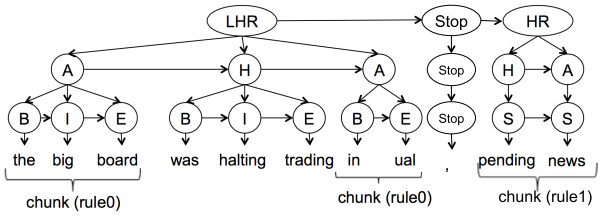


図 8　平坦近似依存構文モデル
**Fig. 8**　Flat approximated dependency grammar model.

以上の対応づけにより構築される HHMM を，**平坦近似依存構文モデル**（FADG）と呼ぶ．FADG は，HHMM の推論アルゴリズムである前向き後向き活性化アルゴリズム [10] によって，教師なし学習および状態の Viterbi 推定を $O(T)$ の計算量で実行することができる．

### 4.3　チャンクの抽出

Viterbi 推定によって得られた状態系列に基づいて，チャンクの抽出を行う．**図 8** に，FADG によって得られる状態系列の例を示す．$Q^2 = H$ は主辞のブロックに対応し，$Q^2 \in \{L, R, Arg\}$ は項のブロックに対応していることから，図 8 では階層 2 の状態に，それぞれ主辞と項を表す H と A のラベルを示している．チャンクは以下の 2 つのルールによって抽出する．

rule0：$Q^2 \in \{L, R, Arg\}$ かつ $Q^3 \neq S$ であるブロック．このルールは，項に対応する 2 語以上のブロックをチャンクとして抽出する．

rule1：$Q_t^3 = S$，$Q_{t+1}^3 = S$，$Q_t^2 = H$ かつ $Q_{t+1}^2 = R$ を満たす部分系列 $w_{t:t+1}$．チャンクの定義から，項が 1 語の場合は主辞を含んだブロックがチャンクといえる．rule1 は，この定義に基づいてチャンクを抽出する．

## 5.　実験

### 5.1　実験方法

提案手法の有効性を確認するため，（1）Ponvert ら [7] の論文に掲載されている結果との精度比較，（2）ドメインの異なるコーパスでの教師ありチャンキング手法との比較，（3）教師なし構文解析モデルである DMV との実行時間の比較の実験を行う．

Ponvert ら [7] の結果と比較するため，同一の条件で実験を行う．実験コーパスには，Penn Treebank の Wall Street Journal を用いる．本コーパスは 24 のセクションに分かれており，学習データとしてセクション 0 から 22 までの単語系列を，テストデータとしてセクション 23 を用いる．Penn Treebank のアノテーションには，名詞句（NP）を表すタグがついているため，名詞句のみを正解チャンクとした場合と，チャンクの定義に従ったすべてのチャンクを正解とする場合の 2 通りで評価を行う．推定したチャンク

表 1　チャンク抽出精度の比較
**Table 1**　Accuracy of unsupervised chunking.

| | Chunking | | | NPS | | |
|---|---|---|---|---|---|---|
| Method | Prec. | Rec. | F | Prec. | Rec. | F |
| IOB-HMM | 53.8 | 62.2 | 57.7 | 47.7 | 65.6 | 55.2 |
| IOB-PRLG | 76.2 | 63.9 | 69.5 | 76.8 | 76.7 | 76.7 |
| FADG rule0 | **78.2** | 63.5 | 70.1 | **79.3** | 76.7 | **78.0** |
| FADG rule0+1 | 72.6 | **71.6** | **72.1** | 65.8 | **77.1** | 71.0 |

は，正解チャンクの区間と完全一致しているもののみを適合とし，精度，再現率，F 値によって評価する．平坦近似依存構文モデルの学習では，前向き後向き活性化アルゴリズムに基づく EM アルゴリズムを 20 iteration で行う．チャンクの抽出は，rule0 のみで抽出した場合と，rule0 と rule1 の両方で抽出した場合の 2 通りで評価を行う．

教師ありチャンキングとの比較では，Wall Street Journal のセクション 0 から 22 を学習データとして構築した条件付き確率場（Conditional Random Field, CRF）との比較を行う．チャンキング対象のデータには Penn Treebank の Brown コーパスを用いる．IOB-PRLG と FADG の学習にはタグが不要なので，チャンキング対象である Brown コーパスを学習データとして用いる．CRF の素性は，Sha ら [9] の手法に従って構築した．CRF の適用には品詞のタグ付けが必要になるが，Wall Street Journal のタグ付きコーパスから学習された最大エントロピー法に基づく tagger を利用した．CRF の実装は CRF++ [*1]，tagger の実装は NLTK [*2] を用いた．

DMV との実行時間の比較では，Wall Street Journal のセクション 23 を学習データとして，EM アルゴリズムの 1 iteration の実行にかかる時間を計測する．IOB-PRLG と FADG については，20 回の計測を行った平均によって評価する．

### 5.2　実験結果

**表 1** に，IOB-HMM，IOB-PRLG，平坦近似依存構文モデル（FADG）の抽出精度の結果を示す．Chunking の列にはすべてのチャンクを正解とした場合，NPS の列には名詞句のみを正解と見なした場合の評価値を示している．rule0 のみでの結果を見ると，FADG の再現率は IOB-PRLG とほとんど変化していないが，精度が向上している．ここから，統語構造の考慮によって主辞に対応するブロックを効果的に抽出結果から除去できているといえる．また，rule0 と rule1 を両方適用した場合の結果では，すべてのチャンクを正解とした場合の再現率が大きく向上している．rule1 で抽出できるチャンクは前置詞句や形容詞句が多いため，名詞句の精度は特に悪化するが，名詞句以外のチャンクも評価に入れた場合には，従来手法では抽出が難しかったチャンクも抽出できていることが分かる．

[*1] http://crfpp.googlecode.com/svn/trunk/doc/index.html
[*2] http://www.nltk.org/

表 2　推定された状態系列の例
**Table 2**　State sequence estimated by FADG and IOB-PRLG.

| | | FADG | | | |
|---|---|---|---|---|---|
| True tag | Input | $d=3$ | $d=2$ | $d=1$ | IOB-PRLG |
| O | at | B | | | O |
| B | 2:43 | I | H | HR | ⌈B |
| I | p.m. | E | | | ⌊I |
| I | edt | S | R | | ⌊I |
| O | , | | | Stop | Stop |
| O | came | S | H | | O |
| B | the | ⌈B | | HR | ⌈B |
| I | sickening | ⌊I | R | | ⌊I |
| I | news | ⌊E | | | ⌊I |
| B | the | ⌈B | | | ⌈B |
| I | big | ⌊I | L | | ⌊I |
| I | board | ⌊E | | LH | ⌊I |
| O | was | B | H | | O |
| O | halting | E | | | ⌈B |
| O | trading | S | L | | ⌊I |
| B | in | ⌈S | H | LHR | O |
| I | ual | ⌊S | R | | O |
| O | – | | | Stop | Stop |
| B | pending | ⌈S | H | HR | ⌈B |
| I | news | ⌊S | R | | ⌊I |
| O | . | | | Stop | Stop |

**表 2** に，推定された状態ラベル系列の例を示す．FADG の $d=3$ の列および IOB-PRLG の列には，タグのラベルとともに抽出されたチャンクの範囲を示しており，⌈ がチャンクの開始，⌊ がチャンクの終了を表す．推定結果を見ると，多くの部分で主辞と項をうまく推定できていることが分かる．たとえば，“the big board was halting” の部分系列では，動詞の “was halting” を主辞，主語の “the big board” を項とする関係を抽出できており，この推定に基づいて動詞のブロックをチャンクから除去している．また，“in ual” の部分系列では，“in” を主辞，“ual” を項としていずれも 1 語のブロックと推定しており，rule1 によりチャンクと推定することに成功している．このような前置詞句は，階層構造を持たない IOB-PRLG のようなオートマトンでは抽出が難しく，教師データを用いずにある程度の統語構造を推定できる提案手法の有効性を示している．

一方，“at 2:43 p.m. edt” の部分系列においては，IOB-PRLG で抽出できているチャンク “2:43 p.m. edt” が，提案手法では抽出に失敗している．本実験では，時刻のような表現も，当該の数字が学習データに現れなければ未知語として扱っているため，チャンク開始の手がかりとして用いることが難しい．IOB-PRLG では，“at” の次の単語には B タグがつきやすい，という知識をパラメータで表現しているために抽出に成功していると考えられる．FADG でも，“at” には主辞の S タグをつけるように学習することで

表 3　ドメインの違いによる精度の比較
**Table 3**　Domain dependency.

| | Chunking | | |
|---|---|---|---|
| Method (Training → Test) | Prec. | Rec. | F |
| IOB-PRLG (Brown → Brown) | 51.0 | 57.1 | 53.9 |
| FADG rule0+1 (Brown → Brown) | 76.0 | 66.2 | 70.8 |
| CRF (WSJ0-22 → WSJ23) | 93.2 | 93.3 | 93.3 |
| CRF (WSJ0-22 → Brown) | 82.9 | 75.4 | 79.0 |

表 4　WSJ23 の学習の 1 iteration にかかる実行時間の比較
**Table 4**　Execution time for learning WSJ23.

| Method | Execution Time (ms) |
|---|---|
| DMV | 57994131.38 (about 16 hours) |
| IOB-PRLG | 151.50 |
| FADG | 1024.05 |

同様の効果があると考えられるが，ここでは未知語が続くために，時刻を主辞のブロックとして推定してしまっている．このようにデータが疎になっている部分系列では，パラメータの数が少ない IOB-PRLG の方が有効な推定を行う場合がある．しかし，提案手法は教師データを必要としないため，学習データの量を増やすことである程度改善することが期待できる．

**表 3** に，教師あり手法とのチャンキング精度の比較結果を示す．教師ありチャンキング手法である CRF は，学習データと同じドメインである Wall Street Journal のチャンキングでは非常に高い精度を達成しているが，ドメインの異なる Brown コーパスでは精度が悪化している．これは，学習データと異なる語彙や表現が含まれていることによる認識誤りによるものと考えられる．また，CRF では前処理として品詞タグ付けを行う必要があるが，適用した tagger も異なるドメインで学習されたものであり，品詞推定のエラーも精度の低下に影響している．

一方，提案手法である FADG は表 1 の結果と同程度の抽出精度を示しており，コーパスのドメイン依存性の点においては教師ありチャンキング手法よりも有利であるといえる．精度自体は CRF の結果を下回っているため，さらなる改善が必要であるが，教師ありチャンキング手法のドメイン依存性による精度低下に対応するアプローチとして今後の発展の可能性が示唆される．

一方，IOB-PRLG の精度は表 1 の結果と比較して悪化している．Brown コーパスには口語体の文章も多く含まれており，“he told ...” や “to cut ...” などの除去するべきフレーズの割合が増加している．IOB-PRLG はこのようなフレーズもチャンクとして抽出するため，主に偽陽判定が増加して精度が悪化したと考えられる．

**表 4** に，Wall Street Journal のセクション 23 の学習の実行時間の比較を示す．小規模な学習データにもかかわらず，DMV は 1 iteration の実行に 16 時間程度要しており，チャンキング手法と比較して極端に計算コストが高いといえる．DMV は系列長の 3 乗のオーダで実行時間が増加することから，長い文章を含むコーパスへの適用は実用的ではない．この点において，教師なしチャンキング手法の優位性があるといえる．

## 6. 結論

本研究では，局所的な依存構造を考慮し，ブロックの主辞と項の関係を認識する教師なしチャンキング手法を提案した．本手法では，平坦近似依存文法モデルを階層型隠れマルコフモデルとして形式化し，系列長に対して線形の計算量の推論アルゴリズムによってブロックの推定と局所依存構造の推定をシームレスに行う．実験では，階層構造を持たない確率オートマトンでは抽出が難しかったチャンクの抽出や，誤って抽出されていたブロックの効果的な除去が可能になることを示した．

平坦近似依存文法モデルは，教師なし構文解析モデルの枠組みに基づいている．このため，DMV の拡張モデルである Extended Valence Grammar [2] に基づいた提案手法の拡張や，DMV の initializer [1] の本手法への適用も考えられるなど，理論的な拡張性の高い枠組みとなっている．また，Ponvert ら [7] はチャンキングモデルを再帰的に適用することで完全な構文解析を行う手法を提案しているが，本手法を応用して完全な構文解析を行う手法も今後の課題である．

**謝辞**　本研究の一部は，JSPS 科研費（課題番号 24800004，25280110，25540159）の助成によって行われた．

### 参考文献

[1] Gimpel, K. and Smith, N.: Concavity and Initialization for Unsupervised Dependency Parsing, *Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.577–581 (2012).
[2] HeaddenIII, W., Johnson, M. and McClosky, D.: Improving Unsupervised Dependency Parsing with Richer Contexts and Smoothing, *Proc. Human Language Technologies: North American Chapter of the Association for Computational Linguistics*, pp.101–109 (2009).
[3] Klein, D. and Manning, C.: Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency, *Proc. Association for Computational Linguistics*, pp.478–485 (2004).
[4] Molina, A. and Pla, F.: Shallow Parsing using Specialized HMMs, *Journal of Machine Learning Research*, Vol.2, pp.595–613 (2002).
[5] Murphy, K. and Paskin, M.: Linear Time Inference in Hierarchical HMMs, *Proc. Neural Information Processing Systems*, pp.833–840 (2001).
[6] Pate, J. and Goldwater, S.: Unsupervised syntactic chunking with acoustic cues: Computational models for prosodic bootstrapping, *Proc. ACL workshop on Cognitive Modeling and Computational Linguistics*, pp.20–29 (2011).

[7] Ponvert, E., Baldridge, J. and Erk, K.: Simple unsupervised grammar induction from raw text with cascaded finite state models, *Proc. Association for Computational Linguistics: Human Language Technologies*, pp.1077–1086 (2011).
[8] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, pp.267–296 (1989).
[9] Sha, F. and Pereira, F.: Shallow Parsing with Conditional Random Fields, *Proc. Human Language Technologies: North American Chapter of the Association for Computational Linguistics*, pp.134–141 (2003).
[10] Wakabayashi, K. and Miura, T.: Forward-Backward Activation Algorithm for Hierarchical Hidden Markov Models, *Proc. Neural Information Processing Systems*, pp.1502–1510 (2012).
[11] 工藤　拓，松本裕治：Support Vector Machine を用いた Chunk 同定，自然言語処理，Vol.43, No.6, pp.1834–1842 (2002).
[12] 北　研二：確率的言語モデル，東京大学出版会 (1999).

**若林 啓**　（正会員）

1984 年生．2012 年法政大学大学院博士課程修了．博士（工学）．同年筑波大学図書館情報メディア系助教．機械学習，自然言語処理の研究に従事．電子情報通信学会，日本データベース学会，ACM 各会員．

（担当編集委員　宮尾 祐介）