自然言語処理 110－12
（1995. 11. 18）

統計によるタグ付きコーパスからの統語規則の獲得

森 信介　　長尾 眞
京都大学工学研究科 電子通信工学専攻
〒606-01 京都市左京区吉田本町
{mori,nagao}@kuee.kyoto-u.ac.jp

あらまし
本論文では、自然言語を文脈自由文法でモデル化し、タグ付コーパスから文法を獲得する方法を提案し、この手法による統語規則の獲得の実験結果を評価する。我々が提案する手法は以下の3つの仮定に基づいている。1) 統語規則の右辺に現れる記号列は環境から独立して高い頻度で現れる。2) 同じ非終端記号から直接導出される品詞列は類似した環境を持つ。3) コーパスの大きさを最も減少させる統語規則が適切である。これらの仮定に基づいてひとまとまりとなる品詞列を見つけ、これを新たな記号で置き換える。これを繰り返すことによって、統語規則の集合 (文法) と統語構造を与えられたコーパスが得られる。
キーワード 文脈自由文法 文法獲得 構文解析 タグ付コーパス N グラム統計

Statistical Grammar Extraction from Tagged Corpora

Shinsuke Mori　　Makoto Nagao
Department of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606-01 Japan
{mori,nagao}@kuee.kyoto-u.ac.jp

Abstract
We describe and evaluate experimentally a method to extract a grammar from a tagged corpus modeling a natural language on context-free language. This method is based on the following three hypotheses. 1) Part-of-speech sequences on the right-hand side of a rewriting rule are less constrained as to what part-of-speech precedes and follows them than non-constituent sequences. 2) Part-of-speech sequences directly derived from the same non-terminal symbol have similar environments. 3) The most suitable set of rewriting rules makes the greatest reduction of the corpus size. Based on these hypotheses, the system finds a set of constituent-like part-of-speech sequences and replaces them with a new symbol. The repetition of these processes brings us a set of rewriting rules, a grammar, and the bracketed corpus.
Key Words CFG, Grammar extraction, Parsing, Tagged corpus, N-gram statistics

1 はじめに

自然言語解析の一般的方法の一つの段階として構文解析がある。構文解析を行なうには、統語規則の集合が必要であるが、これを人手で作成することが困難であるとの理由から、計算機による文法規則の獲得や学習が必要であると考えられるようになった。また、大規模な言語データとそれを処理する計算機資源が比較的容易に入手可能となったことで、大規模コーパスから言語の統計的な性質を得ることが可能になった。このような状況を背景として、我々は Penn Treebank に含まれる品詞情報が付加された英語コーパス Wall Street Journal に対し品詞単位の統計を行ない、その結果を用いた文法規則の獲得について研究を行なった。自然言語のモデルとして、文脈自由文法を用いた。我々が提案する手法は以下の3つの仮定に基づいている。1) 統語規則の右辺に現れる記号列は環境から独立して高い頻度で現れる。2) 同じ非終端記号から直接導出される品詞列は類似した環境を持つ。3) コーパスの大きさを最も減少させる統語規則が適切である。

類似の研究として Fernando Pereira と Yves Schabes [1] は確率文脈自由文法をモデルとし、すべての可能なチョムスキー標準形の書き換え規則の確率をコーパスに付加された統語構造から計算するアルゴリズムを提案している。また、Eric Brill [2] は、まず英語が右分岐の言語であることを根拠にコーパス中のすべての文に対し右線形の統語構造を与え、あらかじめ与えられた正しい統語構造と比較し誤りを最も良く修正する構造変形規則を繰り返し獲得していく手法を提案している。

これらの試みに対して、我々は統語構造情報を用いない、品詞情報のみからの文法規則の獲得の研究を行なった。これは、品詞情報の付加に関しては高い精度のアルゴリズム [3] [4] [5] [6] が提案されていることから修正などの後処理のコストは低いと考えられるが、統語構造情報の付加に関しては高い精度のアルゴリズムが存在しないことから後処理のコストが品詞情報付加に比べて非常に高くなるためである。本論文では、上述した仮定に基づきコーパスを品詞列と見なして統語規則を獲得する手法を提案し、この手法による統語規則の獲得の実験結果を評価する。

2 方法

この章では、まず本研究で用いた自然言語の性質に対する仮定について述べる。次にこの仮定に基づき、自然言語の統語規則をコーパスから獲得し、これを適用する方法を述べる。以下の説明では、終端記号を一般に pos で、非終端記号を syn で表す。終端記号と非終端記号を合わせて tag で表す。また、ボールド文字はそれぞれの記号の長さ1以上の連接を表す。本研究では、コーパスを品詞列とみなすので終端記号は Penn Treebank の品詞記号であるが、非終端記号は任意に導入される。また、以下の説明に現れる Penn Treebank の品詞記号を表1に掲げた。

2.1 統語規則の獲得

統語規則を獲得するためには、その右辺の記号列と左辺の記号を決定する必要がある。以下では右辺の候補となる記号列の獲得と、左辺の記号の導入について述べる。

2.1.1 品詞列の独立性

コーパスから統語規則を獲得するために、統語規則とコーパスの関係について、以下の性質を仮定した。

仮定 1 統語規則の右辺に現れる品詞列は環境から独立して高い頻度で現れる。

ここで、品詞列の環境とはコーパス中における出現の文脈をいう。以下では、統語規則の右辺に現れる品詞列を独立な品詞列と呼ぶ。この仮定の根拠を、例を用いて説明する。独立な品詞列 $\boldsymbol{pos}_a$ と、独立でない品詞列 $\boldsymbol{pos}_x = \boldsymbol{pos}_{x1} \cdot \boldsymbol{pos}_{x2}$ がコーパスに出現したとして、以下の規則が文法に含まれているとする。

$$syn_a \rightarrow \boldsymbol{pos}_a \quad (1)$$

$$syn_b \rightarrow syn_c \cdot syn_d \quad (2)$$

$$syn_c \rightarrow \boldsymbol{pos}_{c'} \cdot \boldsymbol{pos}_{x1} \quad (3)$$

$$syn_d \rightarrow \boldsymbol{pos}_{x2} \cdot \boldsymbol{pos}_{d'} \quad (4)$$

規則 (1) により $\boldsymbol{pos}_a$ は syn_a が出現し得る任意の文脈に出現し得ることがわかる。一方 $\boldsymbol{pos}_x$ は以下のように、規則 (2)(3)(4) により、それぞれ独立に生成された記号列の連接として生成されるので、前後の文脈が制限される。

$$syn_b \stackrel{*}{\Rightarrow} \boldsymbol{pos}_{c'} \cdot \underbrace{\boldsymbol{pos}_{x1} \cdot \boldsymbol{pos}_{x2}}_{\boldsymbol{pos}_x} \cdot \boldsymbol{pos}_{d'} \quad (5)$$

我々は、環境として前後に位置する品詞を選んだ。以上に述べたことから、統語規則の右辺に現れる品詞列の左右に位置する品詞群は、独立でない品詞列の左右に位置する品詞群より多様であると考えられる。左右に位置する品詞の分布は、それぞれ条件付確率とみなすことができ、以下では左右の条件付確率の分布を、それぞれ左確率分布および右確率分布と呼ぶ。品詞群の多様性を表す量として、これらのエントロピーを用いることとした。この値は以下の式を用いて計算される。

$$H_l(\boldsymbol{pos}) = \sum_k P(pos_k \cdot \boldsymbol{pos} \mid \boldsymbol{pos})$$

$$H_r(\boldsymbol{pos}) = \sum_k P(\boldsymbol{pos} \cdot pos_k \mid \boldsymbol{pos})$$

独立性の尺度として、以上に述べたエントロピーに加えて、左右の記号の割合を提案する。ここでいう記号とは、具体的にはピリオドやコンマなどの単語に対応しない品詞を意味する。本研究ではコーパスを文の連接として定義したので、文の先頭に位置する品詞列の左隣と、末尾に位置する品詞列の右隣は、かならず文区切りを表す品詞 "." である。また、独立な品詞列がコンマなどの記号 (以下、

表 1: Penn Treebank POS tagset

1.	CC	Coordinating conjunction	20.	RB	Adverb
2.	CD	Cardinal number	21.	RBR	Adverb, comparative
3.	DT	Determiner	22.	RBS	Adverb, superlative
4.	EX	Existential *there*	23.	RP	Particle
5.	FW	Foreign word	24.	SYM	Symbol
6.	IN	Preposition or subordinating conj.	25.	TO	*to*
7.	JJ	Adjective	26.	UH	Interjection
8.	JJR	Adjective, comparative	27.	VB	Verb, base form
9.	JJS	Adjective, superlative	28.	VBD	Verb, past tense
10.	LS	List item marker	29.	VBG	Verb, gerund or present participle
11.	MD	Modal	30.	VBN	Verb, past participle
12.	NN	Noun, singular or mass	31.	VBP	Verb, non-3rd person singular present
13.	NNS	Noun, plural	32.	VBZ	Verb, 3rd person singular present
14.	NNP	Proper noun, singular	33.	WDT	Wh-determiner
15.	NNPS	Proper noun, plural	34.	WP	Wh-pronoun
16.	PDT	Predeterminer	35.	WP$	Possessive wh-pronoun
17.	POS	Possessive ending	36.	WRB	Wh-adverb
18.	PRP	Personal pronoun	37.	,	Comma
19.	PRP$	Possessive pronoun	38.	.	Sentence-final punctuation

デリミターと呼ぶ) を含むことは少なく、文が独立な品詞列とデリミターの連接として表されると考えられるので、これらの記号に隣接する品詞列は独立である可能性が高い。よって、品詞列の独立性の基準として左右のデリミターの割合を利用することができる。本研究の実験では、文末記号とコンマ以外の記号を含む文を対象としなかったので、デリミターは "," と "." である。

以上、2 つの独立性の基準について説明した。品詞列の抽出を実際に行なう際は、これらの値に閾値を設け、閾値との大小関係によって独立であるか否かを決定した。また、一定の出現頻度がないと統計値が信頼できないため、出現頻度にも制限を設けた。以上をまとめると、現頻度の閾値を f_{min}、エントロピーの閾値を H_{min}、デリミターの条件付き確率の閾値を Pd_{min} として、品詞列 pos が独立である条件は以下の不等式をすべて満たすことである。

1. $f(pos) \geq f_{min}$
2. $H_l(pos) \geq H_{min}$
3. $H_r(pos) \geq H_{min}$
4. $Pd_l(pos) = P(".\cdot pos \mid pos) + P(",\cdot pos \mid pos) \geq Pd_{min}$
5. $Pd_r(pos) = P(pos\cdot ." \mid pos) + P(pos\cdot ," \mid pos) \geq Pd_{min}$

例えば、$f_{min} = 8$, $H_{min} = 3$, $Pd_{min} = 0.05$ とした場合、以上の条件を満たす品詞列は 429 個であった。これらの中で頻度が高い順に 10 個を表 2に掲げる。

表 2: 抽出された品詞列の例

f	H_l	H_r	Pd_l	Pd_r	pos
54724	3.1	3.3	0.16	0.19	NNP
39375	3.6	3.6	0.05	0.27	NNS
23758	3.2	3.5	0.20	0.17	DT·NN
20088	3.3	3.4	0.24	0.25	NNP·NNP
19055	3.1	3.8	0.09	0.06	VBD
18460	4.2	4.0	0.17	0.16	RB
14550	3.7	3.1	0.06	0.17	CD
13327	3.6	3.4	0.06	0.08	VBN
9519	3.8	3.5	0.08	0.34	JJ·NNS
9385	3.2	3.2	0.07	0.22	IN·NNP
9278	3.4	3.5	0.09	0.24	IN·DT·NN
7753	3.5	3.0	0.33	0.05	PRP
7296	3.8	3.5	0.09	0.33	NN·NNS
7487	3.0	3.3	0.15	0.25	DT·JJ·NN

2.2 品詞列の類似性

獲得された品詞列を右辺に置き、左辺に非終端記号を置くことで統語規則を獲得することができる。この際の、左辺の記号を決定するために以下の性質を仮定した。

仮定 2 同じ非終端記号から直接導出される品詞列は類似した環境を持つ。

以下では、この仮定の根拠を、以下の規則が文法に含まれている場合を例として述べる。

$$syn_e \rightarrow pos_{e1}, \quad syn_e \rightarrow pos_{e2}$$

これは、2つの品詞列 pos_{e1} と pos_{e2} の左右に位置することができる品詞はどちらの品詞列の場合も syn_e の左右に導出することができる品詞であることを意味する。そこで大数の法則を用いると、pos_{e1} と pos_{e2} の出現頻度がある程度大きければ、それぞれの左右の品詞列の確率分布は互いに類似するものと考えられる。

本研究では、以上の仮定を用いて品詞列の左右の1品詞の確率分布をベクトルとみなし、2つの品詞列の類似度がこの距離で表されるとして、複数の品詞列に対してクラスタリングを行なった。このクラスタリングは以下の4つ処理で構成される (図1参照)。

1. 独立な品詞列を抽出し、それぞれの間の類似度を計算する。
2. ノードに品詞列が、アークに品詞列間の類似度が対応するグラフを生成する (完全無向グラフ)。
3. ある閾値以下の値を持つアークを消去する。
4. グラフの連結性を調べ、連結成分に分割する。

この結果分割された部分グラフがクラスターに対応し、それぞれのノードが要素に対応する。例として、前項で抽出例として挙げた品詞列に対して、類似度の閾値として $D_{min} = 0.25$ を用いた場合の結果を表3に掲げる。表中の「出現合計」は、各クラスターに与えられる数値で、要素である各品詞列のコーパス中の出現頻度と長さの積の合計である。これを計算する際に、同一クラスターに属する要素の中に一方が他方の部分文字列になる場合があるが、ある品詞列の出現頻度計算する際には同一クラスター内にその部分文字列が含まれる場合はこの頻度を引いて考えることとした。よって、出現合計はクラスターから得られる統語規則をコーパスに適応した場合のコーパスの長さの減少量と見なすことができる[1]。

以上のようにして得られる品詞列のクラスターは、同一の非終端記号に書き換えられると考えられる。これを具体的に説明するために例として表3の一番上にあるクラスターをとる。非終端記号として syn_1 を導入すると、このクラスターから以下の統語規則群が得られる。

$$syn_1 \rightarrow VBD$$
$$syn_1 \rightarrow RB \cdot VBD$$

[1] 各クラスター間の出現合計の大小関係とコーパスの長さの減少量の大小関係は、必ずしも同値ではない。これは、ある同一の品詞列をそれぞれ左端と右端にもつ品詞列が含まれる場合 (例：“DT NN”と“NN NNS”) などがあるためである。しかし、大小関係に逆転が起こる場合はまれであり、これが起こったとしても以後の処理に対する影響は少ないと考える。

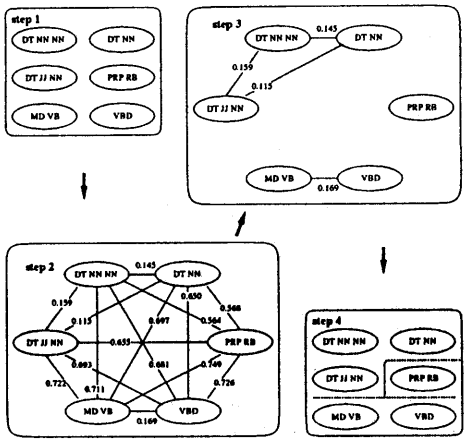

図 1: 品詞列のクラスタリング ($D_{min} = 0.25$)

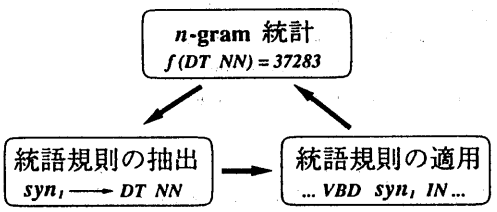

図 2: 処理の概念図

この例では出現合計が 20,260 である。これは、この規則をコーパスに適用すると最大 20,260 個の記号が書き換えの対象となることを意味する。

2.3 抽出と適用の反復

前節で述べた方法で、各クラスターから統語規則群を得ることができる。品詞列であるコーパスに対してこの方法を適用して得られる統語規則は、右辺が終端記号列の統語規則に限られる。しかし、文脈自由文法では右辺に非終端記号を持つ統語規則もあり得る。このような規則を獲得するために、すでに獲得された統語規則を適用した結果得られる非終端記号を含む記号列もコーパスとみなして、統計・獲得・適用の処理を繰り返し行なうことを考えた。この繰り返しは、以下のようにまとめられる (図2参照)。

1. コーパスに対して n-gram 統計をとる。
2. n-gram 統計の結果をもとに統語規則を獲得する。
3. 得られた統語規則をコーパスに適用する。
4. 処理1へ戻る。

以下では、これらの3つの処理を詳しく説明する。

2.3.1 n-gram 統計の計算

n-gram 統計には Nagao と Mori が提案したアルゴリズム [7] を用いた。このアルゴリズムでは、まずコーパスをソートし、その結果を順に見ていくことで n-gram の頻

表 3: クラスタリング結果の例

出現合計	f	H_l	H_r	pos
20260	19055	3.1	3.8	VBD
	1205	3.2	3.8	RB·VBD
86473	23758	3.2	3.5	DT·NN
	7487	3.0	3.3	DT·JJ·NN
	3905	3.0	3.7	DT·NNS
	2137	3.1	3.5	PRP$·NN
	1106	3.2	3.5	DT·NN·IN·DT·NN
	1102	3.1	3.7	DT·JJ·NNS
80317	9278	3.4	3.5	IN·DT·NN
	6553	3.2	3.4	IN·NN
	3722	3.1	3.6	IN·NNS
	3283	3.0	3.5	IN·DT·JJ·NN
	2555	3.2	3.3	IN·JJ·NNS
	1771	3.1	3.3	IN·JJ·NN
	1499	3.6	3.7	IN·DT·NNS
	1326	3.4	3.3	IN·DT·NN·NN

「出現合計」は、各クラスターに与えられる数値で、要素である各品詞列のコーパス中の出現頻度と長さの積の合計である。

度を効率良く計算している。ある n-gram の左右のエントロピーは、その n-gram の出現頻度と $(n+1)$-gram の出現頻度を比べることで計算されるが、我々は Nagao と Mori が提案したアルゴリズムを拡張し、頻度と左右のエントロピーを同時に計算することを可能とした。この拡張によって、エントロピーを求めるための n-gram の出現頻度と $(n+1)$-gram の出現頻度の比較にかかる計算コストを削減できた。この段階での出力はコーパス中に現れるすべての n-gram の頻度と左右のエントロピーである。

2.3.2 統語規則の抽出

すでに説明した方法で n-gram 統計の結果からコーパス中に現れる記号列のクラスターを得ることができる。この結果、クラスターの要素を右辺とし、左辺にある記号を導入することで統語規則を得ることができる。このときクラスターに、1つの非終端記号からなる品詞列が1つ含れる場合は新たな記号を導入する必要はないので、これを左辺の非終端記号とし右辺の品詞列群から取り除く。それ以外の場合は、新たに非終端記号を導入し左辺とする。

以上のようにして獲得された統語規則をコーパスに適用するのであるが、それぞれの統語規則群が互いに関係することがあり、適用順序に問題が生じることがある。例えば、表3の中の2番目のクラスターの最初の要素から、

$$syn_2 \rightarrow DT \cdot NN \tag{6}$$

が得られ、3番目のクラスターの3番目の要素から、

$$syn_3 \rightarrow IN \cdot DT \cdot NN \tag{7}$$

が得られる。ここで、syn_2 と syn_3 は、2番目のクラスターと3番目のクラスターのために導入される非終端記号である。この例を見ると、統語規則 (7) は、統語規則 (6) を用いて、

$$syn_3 \rightarrow IN \cdot syn_2$$

とすることができる。この例のように、それぞれの統語規則群が互いに関係することがあるので、抽出されたすべてのクラスターを統語規則に変換し、これらをコーパスに適用することはできない。このため、唯一のクラスターを選択することにした。次に問題となるのがクラスターの選択基準であるが、この基準を与えるために以下のような第三の仮定を設けた。

仮定 3 コーパスの大きさを最も減少させる統語規則が適切である。

コーパスの大きさの減少量が大きいということは、そのクラスターが大域的構造に関係なく現れる基本的な記号列であるという考えが、この仮定の根拠である。すでに述べたように、コーパスの大きさの減少量は各クラスターに与えられた出現合計の値で表されているので、この仮定は出現合計の値が最大となるクラスターから統語規則を獲得することを意味する。クラスタリングの結果の例として掲げた表3の計算に用いた条件で出現合計が最大となったのは2番目のクラスターであり、従来の文法における名詞句がこれに相当すると考えらる。従来の文法における名詞句は文の基本的な構成要素であり文の大域的構造との関係は少ない。これは上記の推論を傍証する結果である。

後続する処理でこのようにして得られた統語規則をコーパスに適用するが、これを行なうことでコーパスに変化が生じる。この結果、クラスタリングで選択されなかった記号列の確率分布が変化し、記号列の間の類似性に変化が生じる可能性がある。例として、選択されたクラスターに記号列 $DT \cdot NN \cdot NN$ が含まれ、記号列 $DT \cdot NN$ が含まれず、選択されたクラスターに属する記号列の特徴の一つとしてその右に品詞 NN が出現する確率 $P(pos \cdot NN \mid pos)$ が低いことが挙げられる場合を考える。クラスタリングの段階での品詞列 $DT \cdot NN$ の右確率分布の計算では、記号列 $DT \cdot NN \cdot NN$ がひとまとまりとなることが分かっていなかったので、$P(DT \cdot NN \cdot NN \mid DT \cdot NN)$ の値は高くなるが、クラスター抽出の結果 $DT \cdot NN \cdot NN$ がひとまとまりの記号列として書き換えられることが分かったので、記号列 $DT \cdot NN$ の右確率分布の計算においてこれを考慮するべきである。コーパスの中のすべて箇所で記号列 $DT \cdot NN \cdot NN$ が書き換えの対象となるかはこの段階で全く分からないので、これを除いた場合の

$P(DT\cdot NN\cdot NN \mid DT\cdot NN)$ の値を計算することはできない。そこで、コーパスの長さの減少量が最大になる統語規則が適切であるという仮定を用いて、コーパスの中のすべて箇所で記号列 $DT\cdot NN\cdot NN$ がある非終端記号に書き換えられるとし、$P(DT\cdot NN\cdot NN \mid DT\cdot NN)=0$ とした。具体的には、クラスタリングによって得られた統語規則をコーパスに暫定的に適用し、その結果得られるコーパスに対して再度確率分布を計算する手法を用いた。このとき、独立な記号列に類似する記号列は独立であると考えられるので、候補となる記号列の条件は統計結果の信頼性が限界となるある出現頻度 F 以上であることのみとした。この結果得られる記号列の確率分布に対して統語規則の獲得の際に導入された非終端記号の確率分布との距離を計算し、これが閾値 D_{min} 以下となる記号列を新たにクラスターに加えることとした。このとき、この記号列が非終端記号を含む場合などがあり得るので、この結果得られる記号列を以下のように分類し、それぞれに対して異なった処理を行なった。ただし、クラスタリングによって獲得した統語規則を $syn_c \rightarrow tag_i$ $(i=1,2,\ldots,k)$、これをコーパスに適用した結果、syn_c との類似性が条件を満たす記号列を $tag = tag_1\cdot tag_2\cdots tag_n$ とする。

1. $syn_c \notin tag$ (獲得した統語規則の左辺の非終端記号を含まない場合)

   処理：得られた記号列を右辺とする統語規則 $syn_c \rightarrow tag$ を獲得する。

2. $tag = syn_c\cdot tag_1$ (獲得した統語規則の左辺の非終端記号を左端に含み、長さが2である場合)

   処理：syn_c をクラスタリングによって得られた統語規則の右辺で置き換える。この結果得られる記号列の中でコーパス中に存在する記号列を右辺とする統語規則 $syn_c \rightarrow tag_i\cdot tag_2$ $(i=1,2,\ldots,k)$ を獲得する。ただし、右辺の記号列がコーパス中に存在しない場合を除く。

3. $tag = tag_2\cdot syn_c$ (獲得した統語規則の左辺の非終端記号を右端に含み、長さが2である場合)

   処理：syn_c をクラスタリングによって得られた統語規則の右辺で置き換える。この結果得られる記号列の中でコーパス中に存在する記号列を右辺とする統語規則 $syn_c \rightarrow tag_1\cdot tag_i$ $(i=1,2,\ldots,k)$ を獲得する。ただし、右辺の記号列がコーパス中に存在しない場合を除く。

4. 上記以外の場合。

   処理：いかなる統語規則も獲得しない。

以上のようにして更新される統語規則群を再びコーパスに適用し、新たな統語規則が得られなくなるまでこれを繰り返す。最後に獲得された統語規則群の右辺に含まれる単独の非終端記号の数を数え、これが1であった場合、統語規則群の左辺にある暫定的に導入された非終端記号をこの単独の非終端記号で置き換え、この結果生じる左辺と右辺が等しい統語規則を取り除く。以上の処理の結果、同一の非終端記号に書き換えられる記号列に対する統語規則が獲得される。

2.3.3 コーパスへの統語規則の適用

コーパスへの統語規則の適用とは、コーパスの中で統語規則の右辺にある記号列と一致する部分を同一の統語規則の左辺にある記号列で置き換えることである。このとき、複数の統語規則が適用できる場合や、同一の統語規則が部分的に重複する複数の箇所に対して適用可能な場合がある。このような場合は可能な書き換えの中から唯一の書き換えが排他的に選択されるとした。このそれぞれについて順に説明する。

複数の統語規則の競合　複数の統語規則が競合する場合には、右辺がより長い統語規則を優先して適用することとした。これは、統語規則群の選択と同様にコーパスの長さの減少量が最大になる統語規則が適切であるという仮定による。例として、2つの統語規則

$$syn_1 \rightarrow tag_1\cdot tag_2 \tag{8}$$

$$syn_1 \rightarrow tag_1\cdot tag_2\cdot tag_2 \tag{9}$$

を仮定し、コーパスに以下の記号列が存在すると仮定すると、

$$\ldots tag_1\cdot tag_2\cdot tag_2\ldots$$

統語規則 (8) 及び (9) は、両方とも適用可能であるが一方を適用すると他方が適用できなくなるので、どちらか一方の統語規則を選択しなければならない。このような場合は、右辺がより長い統語規則 (9) を優先して適用することとしたので、この例では、以下の記号列が得られる。

$$\ldots syn_1\ldots$$

同一の統語規則の競合　同一の統語規則が重複する複数の箇所に対して適用可能な場合は、単純に左から書き換えを行なうことにした。例えば、統語規則が $syn_1 \rightarrow tag_1\cdot tag_1$ であり、コーパスに以下の記号列が存在すると仮定すると、

$$\ldots tag_1\cdot tag_1\cdot tag_1\ldots$$

どちらの $tag_1\cdot tag_1$ を syn_1 に書き換えるかで競合する。この場合、左側の $tag_1\cdot tag_1$ が書き換えられ、この結果コーパスは以下のようになる。

$$\ldots syn_1\cdot tag_1\ldots$$

左からの書き換えにより、獲得される文法は左分岐の傾向をもつと考えられるが、実験の結果得られた文法は右分岐の傾向を持つもので、英語に対する我々の言語直観と一致する。このことは、書き換えを行なう方向は我々が提案する文法獲得手法に本質的な影響を与えなかったことを意味する。

2.3.4 反復による新たな規則の獲得

得られた統語規則群を適用することでコーパスの状態が変化する。変化したコーパスに対して再び統計をとることで新たな統語規則群が得られる。しかし、この繰り返しはコーパスに含まれる記号数を減らすので、統計結果の信頼性は徐々に減少することに注意しなければならない。つまり、出現頻度が非常に低い記号列の左右のエントロピーや左右の確率分布を用いて独立性や類似性を議論することの危険性は、出現頻度の低下にともなって増大する。この点を考慮して、統語規則が新たに獲得できない場合の処理を以下のように設定した。ここで、f_{min} は独立な記号列を抽出するときの頻度の閾値であり、F は統計結果の信頼性が限界となる出現頻度で、あらかじめ与えられる。

処理 統語規則が新たに獲得できなければ f_{min} を 0.8 倍する。この結果

1. $f_{min} \geq F$ であれば、統語規則の獲得を実行する。
2. $f_{min} < F$ であれば、終了する。

3 評価

前節で述べた方法を用いて、実際に統語規則の獲得を、表 4 に示される 2 組のパラメータを用いて行なった。実験の結果得られる出力は、コーパスの解析結果、すなわちコーパスの最終状態と統語規則の集合である。以下では、実験に用いたコーパスについて述べた後、それぞれの結果を提示し、これらの結果に対する評価と考察を行なう。

表 4: パラメータ

	f_{min}	H_{min}	Pd_{min}	D_{max}	F
実験 1	2000	3.0	0.05	0.25	50
実験 2	2000	3.0	0.05	0.20	50

3.1 用いたコーパス

実験に用いたコーパスは Penn Treebank に含まれる Wall Street Journal である。これは、実在する雑誌の文章を計算機可読にしたコーパスであり、括弧やその他の記号を含む文もある。人間がこれらの記号を含む文章を読むときは記号の対称性などの視覚的情報も用いると考えられるが、本研究ではこれを対象外としているので、実験を行なう前に文末記号とコンマ以外の記号を含む文を取り除いた。よってコーパス中に含まれるのは、単語に対応する品詞タグと文末記号とコンマに対応する品詞タグである (表 1 参照)。以上の処理の結果、コーパスに含まれる文の数は 24,678 であり、品詞の数は 549,247 であった。1 文あたりの品詞数は 22.3 である。

3.2 解析結果の評価

ここでは、構文解析の結果とみなすことができる、コーパスの最終状態について述べる。解析結果の例を図 3 に示す。評価基準として、Crossing Parenthesis Accuracy [8] を用いることにした。これは、出力された構造を示す括弧のうち、Penn Treebank に人手で与えられた統語構造を表す括弧と交差しない括弧の割合である。表 5 はそれぞれの実験の Crossing Parenthesis Accuracy を示す。表中の「右線形」とは、文末記号を除いた記号列に対して最も右に位置する 2 つの記号を 1 つの記号で書き換えることを繰り返して得られる構造である。同様に「左線形」とは、これを左から順に行なうことで得られる構造である。

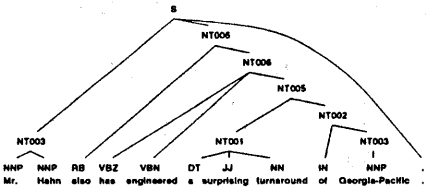

図 3: 解析結果の例 (実験 2)

表 5: 実験結果

	規則数	非終端記号数	C.P.A.
右線形	—	—	56.3%
左線形	—	—	24.3%
実験 1	956	19	73.7%
実験 2	594	25	74.8%

C.P.A. = Crossing Parenthesis Accuracy

それぞれの実験の Crossing Parenthesis Accuracy は、コーパスの統語構造を用いた文法推論 [1] [2] の結果 (90% 前後) より少し低い程度であり、我々が提案する手法を用いて、統語構造を用いない文法獲得が可能であることを意味すると考えられる。

3.3 統語規則の評価

本研究の目的は統語規則の獲得であるが、統語規則を定量的に評価することは困難である。従って、以下では一般的な英文法との比較を定性的に行なう。

実験 2 によって獲得された文法の中から比較的単純な統語規則を表 6 に掲げる。表中の $NT001$ は名詞句を表し、$NT002$ は前置詞句を表すと考えられる。また左辺が $NT006$ である統語規則は、右辺に名詞句や前置詞句をさまざまな組合せで含む。これらは、動詞の格フレームを表すと考えられ、Brent と Berwick [9] や Brent [10] や Manning [11] の動詞の格フレーム獲得の研究においてあらかじめ仮定されていた動詞の格フレームのタイプが自動的に獲得可能であることを示す結果である。また、規則の多くが右分岐の傾向を持つ点にも注意する必要がある。

しかしながら、獲得された規則の中には、不適切である

と思われる規則も含まれていた。例えば統語規則

$$NT002 \rightarrow IN \cdot NNS \cdot NN$$

は、後ろの2つの品詞を $NT001$ (名詞句) に書き換える統語規則

$$NT001 \rightarrow NNS \cdot NN$$

に置き換える方がより適切であると考えられる。

このように、獲得した文法の中には再考を要する統語規則も含まれるが、導入された非終端記号の多くは我々の言語直観と対照して解釈することが可能であった。

表 6: 実験 2 で得られた規則の一部

NT001	→	PRP$ · NNS · NNS
NT001	→	NNP · NNP · POS · NN
NT001	→	DT · JJ · NN
	⋮	
NT002	→	TO · NT001
NT002	→	IN · NT001
NT002	→	IN · NNS · NN
	⋮	
NT003	→	NNP · NNP · NNP
NT003	→	NT003 · CC · NT003
	⋮	
NT004	→	NN · NN
NT004	→	JJ · NN · NN
	⋮	
NT005	→	NT001 · NT002
NT005	→	NT001 · VBN · NT002
NT005	→	VBN · NT002
	⋮	
NT006	→	VBZ · NT002 · NT005
NT006	→	VBZ · NT002
NT006	→	VBZ · NT001 · NT001
NT006	→	VBZ · NT001
	⋮	

4 おわりに

本論文では品詞情報が付加されたコーパスから統計的手法を用いて自然言語の統語規則を獲得する方法について述べた。我々が提案する統語規則獲得の手法の特徴は以下の3点である。

1. 自然言語を文脈自由文法でモデル化した。
2. 付加するコストが低い品詞情報のみを用いる。
3. 統計的手法によるためコーパスに含まれる誤りの影響を受けにくい。

獲得された文法には、我々の言語直観に適合する統語規則が多数含まれていた。従来の統語情報を用いた文法推定に対し、品詞情報のみを用いて統語規則を獲得することでも、比較的高い精度が得られた。これは、作成コストがより低い品詞情報からの統語規則獲得が有効であることを意味する。

参考文献

[1] Fernando Pereira and Yves Schabes. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of ACL*, 1992.

[2] Eric Brill. Automatic grammar induction and parsing free text: A transformation-based approach. In *Proceedings of the 31st Annual Meeting of ACL*, 1993.

[3] Kenneth Ward Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on ANLP*, 1988.

[4] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on ANLP*, 1992.

[5] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on ANLP*, 1992.

[6] Eric Brill. Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelveth AAAI*, 1994.

[7] Makoto Nagao and Shinsuke Mori. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. In *Proceedings of the 15th Coling*, 1994.

[8] E. Black et. al., A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 1991.

[9] Michael R. Brent and Robert C. Berwick. Automatic acquisition of subcategorization frames from tagged text. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 1991.

[10] Michael R. Brent. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of ACL*, 1991.

[11] Chiristopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of ACL*, 1993.