自然言語処理 114－4
(1996. 7. 18)

話し言葉用文法のBigramの追加による自動学習法

大谷耕嗣　中川聖一

豊橋技術科学大学 情報工学系
〒441 豊橋市天伯町字雲雀ヶ丘1-1
Tel. (0532) 44-6777
E-mail {otani, nakagawa}@slp.tutics.tut.ac.jp

[あらまし] 本稿では文のカバー率を改善するための文法規則の自動学習法について述べる。この方法は文法規則が登録されていないために解析できない文を解析することを可能にする。システムに入力された文が文法規則が不備なために受理できない時、システムがこの入力文を使って規則の学習することにより文のカバー率を改善する。

未登録単語の登録法以外に、文法の学習の3つの方法を比較検討する。1つは、文脈自由文法の規則の追加学習法、2つめは単語ペアの追加学習法、3つめは非終端記号のペアの追加学習法である。以上の方法について未登録単語を登録した場合としない場合についての生成規則を学習した後で文のカバー率とパープレキシティ、単語ペアを学習した時には確率を与えた時のバイグラムのパープレキシティについて評価を行なった。

[キーワード] 文法の学習, CFG, 単語ペア, 非終端記号のペア, バイグラム, カバー率, パープレキシティ

An Automatic Learning Method of Grammar Rules for Spontaneous Speech by Adding Bigram

Koji OTANI and Seiichi NAKAGAWA

Department of Information and Computer Sciences
Toyohashi University of Technology
Tenpaku-cho, Toyohashi, 441, Japan
Tel. (0532) 44-6777
Email {otani, nakagawa}@slp.tutics.tut.ac.jp

[abstract] In this paper, we describe an automatic learning method of the grammar rules for improving coverage of the sentences. It is possible to analyze the sentences that can not be accepted by preregistered production rules. When a sentence which is not accepted by production rules is inputted to the system, the system learns rules using this input sentence. As a result, the coverage of the sentence can be improved.

Except for regisering unknown words, there are three methods for learning the grammar. One is to add new production rules to the original CFG. The second is to add word pairs. The last is to add nonterminal symbol pairs. In the registering unknon words or not, we evaluate the coverage of sentences and perplexity using above methods. And we evaluate perplexity using bigram instead of word pairs.

[keyword] learning of grammar, CFG, word pair, nonterminal symbol pair, bigram, coverage, perplexity

1 はじめに

自然言語理解システムの目的の1つに人間・機械間のコミュニケーションの方法としての人間にやさしいインターフェイスの開発があげられる。本研究室では、"富士山のための観光案内システム"のタスクで自然発話を使った対話システムの開発を行なっている [1]。

対話システムはユーザが発話した言葉を受理するための文法を使って発話した文を認識する。しかし、もしユーザがシステムの文法で受理できない文を話した時にはシステムはその文を認識することができない。この原因による認識間違いを減らすために、ユーザがシステムの文法で受理できない文を発話した時に、その文を使って、システムの文法に登録されていない規則の学習を行ない、これらの文を受理できるようにするシステムの開発を行なってきた [2][3][4]。その結果、新しい入力文に対するカバーレイジの改善ができるようになった。

文法獲得の過去の研究で、大量のテキストデータからInside-Outside アルゴリズムで文脈自由文法を学習する方法が検討されているが [5]、非終端記号が増えると学習は非常に難しい。効率良い学習のためには何らかの初期文法が必要である。中澤らはシステムに入力された例文の推測された導出木と現在の文法構造の差から生成規則の追加・削除を行なうことにより、文脈自由文法 (CFG) を効率的に学習する方法を調べている [6]。白井らは、構文構造付きコーパスの内部ノードに非終端記号を与えて確率文脈自由文法を抽出し、その文法を改善することにより適用範囲の広い文法を抽出している [7]。Brill は、句構造の大変ナイーブな状態の知識から始めるアルゴリズムの研究を行なっている [8]。これは、トレーニングコーパスで与えられる解析構造を示す括弧と現在の状態の括弧の結果を比較することを繰り返すことによって、システムがエラーを減らすことができる簡単で構造的な変換のセットを学習する。Miller らは文法内のローカルな文脈的な情報を使うのと、確率付き文脈自由文法の導入の効果について調べている [9]。Wright らは、ロバストなパーザのために CFG とバイグラムの併用を提案している [10]。しかし、それらは統合されておらず並列に実行される。Samuelsson は、解析木のノードに対してのエントロピーをしきい値として使うことによって文法の統合を行なう研究を行なっている [11]。また最近では竹澤らが、ポーズ情報で区切られた区間を部分木で表現し、部分木出力による音声認識実験に対して、前終端記号バイグラムを利用した再順序付けを行なう方法を提案している [12]。これは我々が提案した方法 [3] とほぼ類似な考え方である。

本システムは文法を学習するために3つの方法から構成されている。未登録単語を登録する方法以外に、1つは、規則の不備のための CFG の生成規則を登録する方法である。2つめは接続可能な単語ペア又は単語クラスペアを登録する方法である。単語クラスとは意味論的によく似た単語のセット（例えば、湖名：河口湖, 山中湖, 精進湖等）のことである。最後に今回新たに行なった方法は接続可能な非終端記号のペアを登録する方法である。生成規則の登録は、ボトムアップのパーザを使って新しい適当な単語又は非終端記号のペアを登録する方法である。

未登録単語の登録については [4]、"富士山観光案内システム"を使って評価した結果、この方法は未登録単語を登録するための条件があまりに厳しいためにまだ完全でないことがわかった。生成規則の登録について、3つの登録方法を比べたところ、単語クラスペアがカバーレイジとパープレキシティのバランスが一番良いことが分かった。また、単語（単語クラス）ペアに確率を付与してパープレキシティを抑えることを試み、その有効性を示した。

2節でシステムの認識エラーの原因について、3節で生成規則の登録のための文法登録のアルゴリズムについて、4節で確率の導入について、5節で本手法の評価結果について述べる。

2 システムの認識エラーの原因

本研究室のタスク"富士山観光案内"のための対話システムでの認識エラーは以下の原因によって生じる。

- 文法部のエラー
- 認識部のエラー

前者のエラーはユーザが受理できない未登録単語（語彙外の単語）を含む文を話した時か生成規則で受理できない文を話す（文法外）時のエラーである（もちろん、これはシステムの語彙と文法規則がまだ完全ではないともいえる）。後者のエラーはユーザが文法で受理できる文を話したにもかかわらずシステムの音声認識部がを正しく認識できなかった場合のエラーである。

以上の2つの問題を解決することによって認識エラーを減らすことができる。本稿では、前者のエラーを減らすために受理できなかった文を学習に使うことによって文法の新しい規則を半自動的または自動的に登録する方法について述べる。

前者のエラーには3つの場合がある。それらは以下の登録の欠落である。

- 単語
- 生成規則
- 単語と生成規則

システムは1番目と2番目のエラーについての登録を行なう。3番目のエラーに対しては現在は対応していないが、未登録単語を人間によって完全に登録したバージョンを作成して完全に未登録単語をなくした場合での評価を行なった。

3 生成規則登録のアルゴリズム

本節では例文を使った文法規則を登録するアルゴリズムについて述べる。生成規則の登録は文法の学習に3つの方法を使う。1つはボトムアップパーザを使い CFG の新しい適当な生成規則を作りこれらの新しい規則を登録するもので、2つめの方法は部分的にパーズされたストリングの間の単語のペアまたは単語クラスのペアを登録するもので、最後の方法は部分的にパーズされたストリングのそのストリングに対する非終端記号のペアを登録するものである。

3.1 CFG での規則生成

CFG の生成規則の半自動的な登録方法は既に報告している [4] ので、ここでは簡単に説明しておく。CFG の生成規則登録の要点を以下にまとめた。

1. ボトムアップパーザを使い入力文の部分解析木を作る。
2. 入力文の全てをカバーする部分解析木の組合せのうち組合せの数が最小になる組合せを選ぶ。
3. 先の部分解析木の組合せで、適当な一般性を与えるために 3 つの登録方法を使い分ける。
4. 先の登録方法でいくつかの候補があるのなら、ユーザが適当な候補を判断するためにシステムがその候補を使っていくつかの例文を作りユーザにふさわしいものを選んでもらう。

以上の方法により CFG の生成規則の登録を行なう。

3.2 単語ペアを使った規則の追加

単語（単語クラス）ペアを使うことによる生成規則の不備により解析できない文のための規則の登録方法について述べる。この登録方法は CFG 規則の補助として単語（単語クラス）対制約を登録する方法で、以下の手順で自動的に登録を行なう。

1. システムが入力文の解析に失敗した時に、ボトムアップのパーザを使って入力文の部分解析木を作る。
2. 入力文をカバーできる部分解析木の組合せで、木の組合せの数が最小となる組合せを見つける。
3. 部分解析木の組合せの各解析木に隣接する単語（単語クラス）のペアを登録する。つまり、図 1の様に最小の部分解析木の数が 2 個なら、図中の部分解析木 P_1の最後の単語 w_1と、P_2の最初の単語 w_2の単語（単語クラス）のペア (w_1, w_2) の登録を行なう。

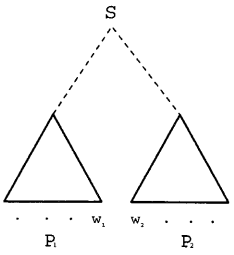

図 1: 単語対制約の登録

この方法で登録した単語（単語クラス）ペアは部分解析木を繋ぐ制約として使用する。但し、実際には、CFG とペアを独立に用いて後続単語の予測を行なう。つまり、CFG 規則と単語ペアは文脈に関係なくいつも使える。

一方、CFG を使わずに入力文から接続可能な単語のペアを獲得して、その単語ペアだけで文の解析をする方法も行なった。

3.3 非終端記号のペアを使った規則の追加

単語（単語クラス）のペアの代わりに、より接続能力の強い非終端記号のペアの登録方法について述べる。この方法により単語ペアの方法よりカバー率の改善が期待される。登録方法は単語（単語クラス）ペアの場合とほぼ同じである。また登録する部分解析木の非終端記号の組合せについては、一番トップにある記号のペアと一番ボトムにある記号のペアを登録した場合について調べた。但し、トップの記号より下位の非終端記号のペアは上位のペアに単語予測に関しては完全に包含される。

1. システムが入力文の解析に失敗した時に、ボトムアップのパーザを使って入力文の部分解析木を作る。
2. 入力文をカバーできる部分解析木の組合せで、木の組合せの数が最小となる組合せを見つける。
3. 部分解析木の組合せの各解析木のペアを登録する。つまり、図 1の様に最小の部分解析木の数が 2 個なら、図中の部分解析木のペア (P_1, P_2) の登録を行なう。第 3.1節の方法と異なる点は、CFG の規則登録と違って任意の場所で使用が可能なペアを登録するが、規則の登録方法と疑似的に等価な関係である。ペアによって予測される非終端記号 P_2より下位の規則の展開に関しては CFG の規則を使う。

この方法も単語（単語クラス）ペアを登録した時と同様に非終端記号のペアを部分解析木を繋ぐ制約として使用する。但し、実際には、CFG とペアを独立に用いて後続単語の予測を行なう。つまり、CFG 規則と非終端記号ペアは文脈に関係なくいつも使える。

4 確率（バイグラム）の導入

単語（単語クラス）ペアのパープレキシティを抑えるために確率を導入した。単語又は単語クラスのバイグラムの推定には学習セットとして使用している 914 文を使用して任意の場所の単語のペア又は単語クラスのペアののの出現頻度からバイグラムの確率を推定している。つまり入力文（単語列）に対して、単語 W_i の出現頻度 $C(W_i)$ と単語の 2 つ組の出現頻度 $C(W_{i-1}, W_i)$ を数え上げる。この時のバイグラム確率は

$$P(W_i|W_{i-1}) = \frac{C(W_{i-1}, W_i)}{C(W_i)} \quad (1)$$

で求まる。単語クラスバイグラムについても同様な式を用いて求めている。ただ、全てのデータを学習した場合でも単語ペアの数は 1250 ぐらいにしかならず、データの不足を補うため単語ペア及び単語クラスペアは学習データの全ての場所を使って求めている。

また頻度が 0 になるペアには頻度を 1/10 としてスムージングした実験も行なったが実験結果にあまり違いはで

なかった（この本稿ではスムージングを行なわない結果だけ報告する）。

バイグラムとCFGにより並列で解析する場合のパープレキシティは以下の2通りの方法で求めた。

方法1 該当単語の予測単語数（分岐数）を、CFGで予測される単語数に$P(W_i|P_{i-1})$の逆数を足したものとし、その相乗平均としてパープレキシティを求める。但し、バイグラムの適用は、CFGの非終端記号が還元された直後だけである。

方法2 該当単語の予測確率を、CFGで予測される確率を予測単語数の逆数とし（該当単語が予測されない場合は確率=0）、これとバイグラムによる予測確率との平均値とした。つまり

$$P(W_i|CFG, bigram) = \lambda P(W_i|CFG) + (1-\lambda)P(W_i|W_{i-1}) \quad (2)$$
(実験では$\lambda = 0.5$)

この逆数をとることによって該当単語の予測単語数（分岐数）とし、これらの相乗平均で求める。但し、バイグラムは任意の場所で適用する。

単語クラスのバイグラムによる単語の確率は、予測確率をその単語クラスに含まれている単語数で割ることによって求まる。また、方法2の場合で、CFGで該当単語が予測されず、バイグラムによる予測確率が0の時には、バイグラムの確率の代わりにユニグラムで求めた確率を使って代用した。

5 評価

5.1 評価条件

実験には" 富士山観光案内" の対話システムの文法を使った。オリジナルな初期文法の詳細は以下の通りである。なお、パープレキシティは学習データ（set 1）でオリジナルな元の文法で解析できた39文を使って求めている。

[初期文法]

- 終端記号数（単語数, 語彙のサイズ） - 241
- 生成規則数 - 393
- 非終端記号数 -137
- 前終端記号数（単語クラスのサイズ） - 110
- **パープレキシティ - 76.5**

[未登録単語登録後]

- 終端記号数（単語数, 語彙のサイズ） - 419
- 生成規則数 - 393
- 非終端記号数 -137
- 前終端記号数（単語クラスのサイズ） - 198
- **パープレキシティ - 89.8**

学習と評価で使われたデータは計1020文を使用した。これらの内106文を評価に、残りの914文を学習に使った。データは発話者にあらかじめ使える単語（名詞と動詞）を教えた条件で集められたものである。表1は学習セットの詳細を示している。表中の" 受理可能文" は学習前のオリジナルな初期文法で受理された文の数を意味している。" 解析失敗の原因" はエラーのタイプを意味している。" 単語" と "規則" は単語又は生成規則の不備により受理できなかった文の数を示している。" 単語 & 規則" は単語と生成規則がともに不備であり、初期文法での学習では対象外とした文の数を示している。

表1: 学習・評価セットの詳細

	受理	解析失敗の原因			合計
	可能文	単語	規則	単語 & 規則	
set 1	39	0	35	32	106
set 2	51	5	32	18	106
set 3	58	4	22	16	100
set 4	31	5	21	43	100
set 5	41	6	19	34	100
set 6	50	3	25	22	100
set 7	43	5	24	28	100
set 8	57	4	26	13	100
set 9	38	7	25	30	100
set 10	50	7	19	32	108
合計	458	46	248	268	1020

今回学習・評価データで現れる未登録単語をあらかじめ登録した実験も行なった。これにより初期文法では対象外であった" 単語 & 規則" の欄の文についても学習の対象となった（それでもなお、評価セットのうち8文が学習セットに出てこない単語が使われているため評価の対象外となっている）。表中のsetのうちset1を評価に、その他のsetを学習に使った。受理可能になる文数は**初期文法での学習の場合39文＋35文＝74文、未登録単語を登録した場合では98文が本研究の上限となる。**

また評価規準としてカバーレイジとパープレキシティを用いた。カバーレイジは評価セットの解析対象文のうち何文が解析できたかを示す。パープレキシティの定義は、2のエントロピー乗で計算される言語の複雑度を示す規準の一つであるが、ここではテスト文集合に対するパープレキシティ（テストセットパープレキシティ）を求めている。

なお、規則や単語（単語クラス）の追加によってカバーレイジは増加するが、確率を使用しない場合は、もとのCFGを含んでいるのでパープレキシティも必ず増加する。

5.2 CFGでの規則登録結果

表2はCFGの規則登録による文法の学習結果についてまとめたものである。" カバー数" はset 1のうち解析できた文の数を示している" 規則数" は登録した規則の数

を示している。**"perp."**はパープレキィティを示し、学習セットのうち受理可能になった文を使って求めている。**"perp2."** はオリジナルな初期文法で解析できた39文を使って求めたパープレキシティである。以下の表の表記も特別の指定がない限りこれに準じている。

CFGを使った生成規則の登録の結果では、システムは生成規則の不備により受理できなかった文の半分が受理できるようになり、受理可能な文のカバーレイジは39文から58文に増加した。しかし、パープレキシティが学習前に比べて約1.5倍以上になってしまった。また、解析に時間がかかりすぎるのも問題である。

表2: CFGの規則登録による文法の学習結果

学習set	カバー数	perp.	perp2.	規則数
set2	52	112.7	110.6	30
set2-3	54	118.1	113.6	45
set2-4	58	125.7	123.3	65
set2-5	58	130.2	126.7	78
set2-6	58	131.7	127.8	98
set2-7	—	—.–	—.–	114
set2-8	—	—.–	—.–	135
set2-9	—	—.–	—.–	154
set2-10	—	—.–	—.–	170

5.3 単語／単語クラス／非終端記号のペアを使った規則登録結果

表3は単語（括弧内は単語クラス）ペアについての学習結果を示している。表中の" ペア数" は登録された単語（単語クラス）のペアの異なり数を示している。表4は学習セットで出現した未登録単語を全て登録した状態での単語（括弧内は単語クラス）ペアについての学習結果を示している。図2、3はカバーレイジについて、図4、5はパープレキシティについてグラフにまとめたものである。

単語クラスのペアによる学習では、最終的に評価の対象となるセット1の74文中59文まで受理可能になった。また、パープレキシティの増加がそれ程でもなくカバーレイジの改善もCFGとほとんど変わらない。よって単語クラスペアの効果はCFGを使った登録よりも良いと言える[4]。

学習セット内の未登録単語を登録した場合の結果は、解析対象文が98文になったこともあるがカバー数が77文(単語クラス)まで増えた。それに伴いパープレキシティがかなり増加しているが、これは未登録単語を登録したことにより単語数が約1.7倍になっている点を考慮すれば、パープレキシティの増加は十分に抑えられていると言える。一方、表5は初期文法のCFGを全く使わない単語ペアと単語クラスペアだけでの学習法による結果を示している（括弧内が単語クラス）。表中の" バイグラム" は確率を導入した結果を示している。単語ペアだけの場合は、学習セットがまだ不十分ではあるが、他の方法と比べてまずまずの結果が得られた。また、単語クラスペアだけによる結果はCFGと単語クラスペアを組み合わせた結果よりも良い結果が得られた。単語クラスによって単語ペアでの学習の場合の学習セットの不足を補うことが出来きることが確かめられた。

表6は非終端記号ペア（トップおよびボトムのペア）の学習の結果を示している。非終端記号のペアの学習によるカバー率の改善はトップのペアを用いた時にはほぼ完全といっても良い結果となった。しかし、パープレキシティが170以上になってしまっている。240単語ぐらいの文法でパープレキシティが170では毎回全単語の7割ぐらいを予測していると考えることもでき、音声認識への応用を考えるとこの非終端記号のトップのペアを使う方法は良くないという結論になる。ボトムのペアも単語ペアと比べてカバー数の割にパープレキシティが大きくなり過ぎている。よって非終端記号のペアを使う方法は良くないと言える。

以上のことから、未登録単語を既に単語クラスに登録してあるのなら単語クラスペアのみによる学習方法が最も良い結果といえる。しかし大語彙の文法を使うならこの方法もよいが、単語数が不完全な文法での学習を考えた時には、単語ペアとCFGを組み合わせる方法のうち単語クラスペアを使う方法が良いと言える。

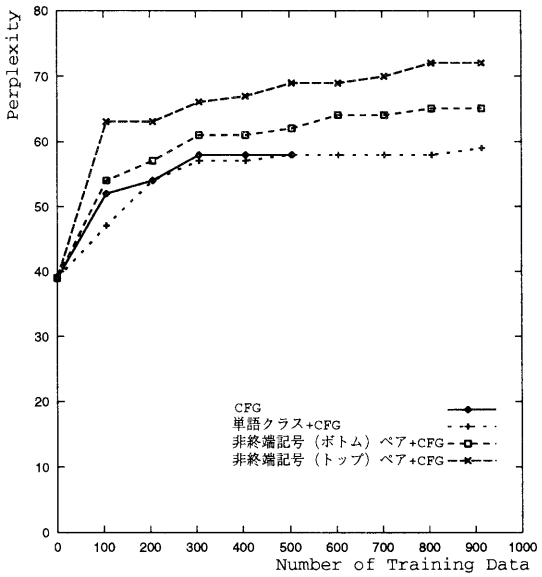

図2: 学習によるカバーレイジの変化
（未登録単語登録前）

5.4 単語／単語クラスのバイグラムによる効果

表7はペアの代わりに確率を使って単語（単語クラス）バイグラムを学習した場合のパープレキシティの結果を示している（注：カバーレイジは単語（単語クラス）ペ

表 3: 単語ペア＋CFG による文法の学習
（括弧内は単語クラスペアの登録）

学習 set	カバー数	perp.	perp2	ペア数
set2	45(47)	83.4(84.1)	84.4(85.0)	60(67)
set2-3	50(54)	91.7(92.7)	95.7(97.1)	83(94)
set2-4	51(57)	91.9(98.0)	95.4(100.1)	103(127)
set2-5	51(57)	92.2(100.7)	95.9(102.7)	123(149)
set2-6	52(58)	92.6(102.1)	95.9(103.6)	144(170)
set2-7	53(58)	92.8(102.5)	95.7(104.1)	164(186)
set2-8	53(58)	93.3(103.1)	96.1(105.8)	183(200)
set2-9	53(58)	93.6(105.5)	96.5(107.7)	206(228)
set2-10	53(59)	93.8(105.8)	96.7(107.8)	216(233)

表 4: 単語ペア＋CFG による文法の学習（419 単語）
(未登録単語をすでに登録した場合)

学習 set	カバー数	perp.	perp2.	ペア数
set2	48(57)	109.3(117.4)	109.1(112.9)	125(147)
set2-3	55(64)	115.2(120.4)	110.4(118.1)	188(212)
set2-4	57(68)	115.9(126.5)	112.2(126.6)	326(351)
set2-5	58(69)	118.3(132.3)	115.7(136.2)	415(437)
set2-6	61(72)	119.4(134.8)	121.1(139.6)	475(491)
set2-7	63(74)	118.4(136.2)	121.7(142.3)	546(544)
set2-8	63(74)	118.6(136.4)	121.8(142.4)	574(561)
set2-9	63(77)	119.3(135.9)	122.5(143.6)	645(624)
set2-10	63(77)	120.0(137.1)	123.1(144.6)	689(655)

表 5: 単語（単語クラス）ペアの学習結果（419 単語）
（括弧内は単語クラスペアおよび
バイグラムのパープレキシティ）

	カバー数	ペア (perp.)	バイグラム	ペア数
set2	12(33)	6.9(21.6)	5.5(16.9)	391(313)
set2-3	21(50)	8.9(29.0)	5.8(19.8)	598(442)
set2-4	30(54)	10.5(35.0)	6.5(21.2)	837(621)
set2-5	39(58)	12.8(38.9)	7.0(21.2)	1024(747)
set2-6	42(61)	13.8(42.5)	7.0(21.6)	1152(820)
set2-7	45(67)	14.5(46.3)	7.1(24.6)	1292(893)
set2-8	48(68)	14.9(46.8)	7.2(24.8)	1366(924)
set2-9	51(71)	16.5(48.6)	7.5(24.8)	1491(996)
set2-10	52(73)	17.3(50.6)	7.8(25.2)	1583(1048)

表 6: 非終端記号+CFG のペアの学習結果
（括弧外ボトム、括弧内トップのペア）

学習 set	カバー数	perp.	perp2.	ペア数
set2	54(63)	122.6(162.1)	124.5(160.0)	147(546)
set2-3	57(63)	125.2(166.0)	127.0(163.2)	199(667)
set2-4	61(66)	131.2(167.6)	131.7(164.0)	246(777)
set2-5	61(67)	133.5(169.4)	134.5(165.2)	290(848)
set2-6	62(69)	134.8(170.0)	135.3(165.9)	341(972)
set2-7	64(69)	140.2(170.1)	143.2(165.9)	404(1077)
set2-8	64(70)	141.4(169.6)	144.5(166.0)	464(1191)
set2-9	65(72)	142.4(174.4)	146.2(170.8)	522(1271)
set2-10	65(72)	142.6(174.4)	146.5(170.8)	549(1298)

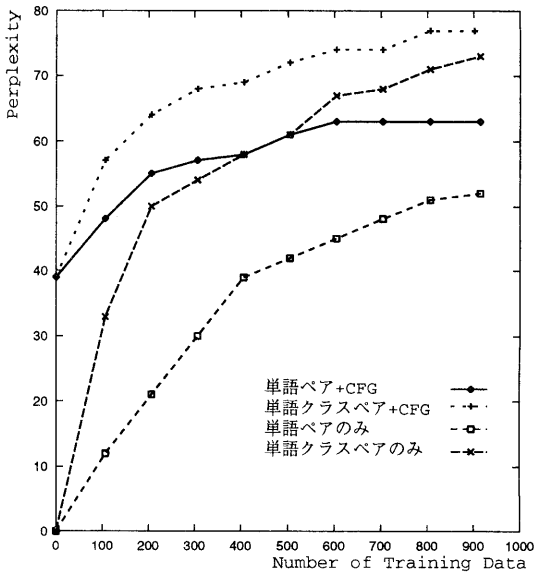

図 3: 学習によるカバーレイジの変化
（未登録単語登録後：419 単語）

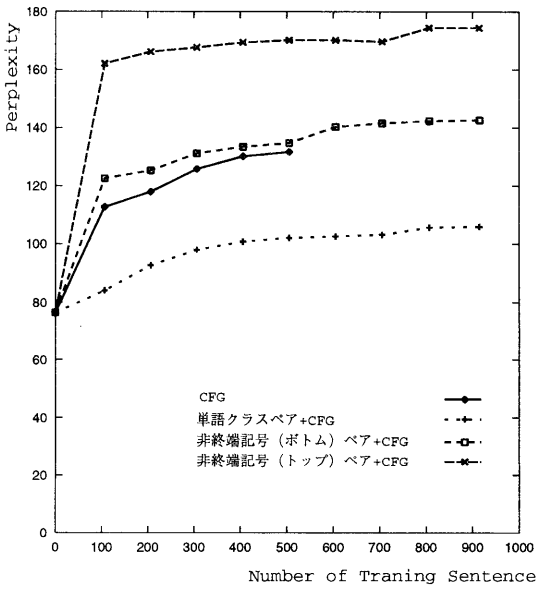

図 4: 学習によるパープレキシティの変化
（未登録単語登録前：241 単語）

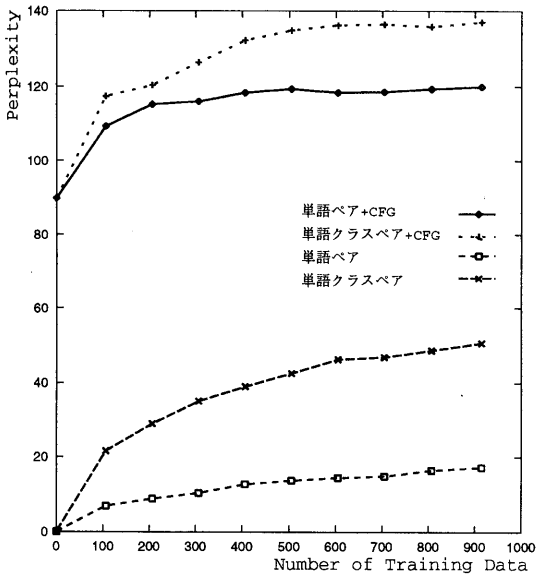

図 5: ペアの登録によるパープレキシティの変化
（未登録単語登録後：419 単語）

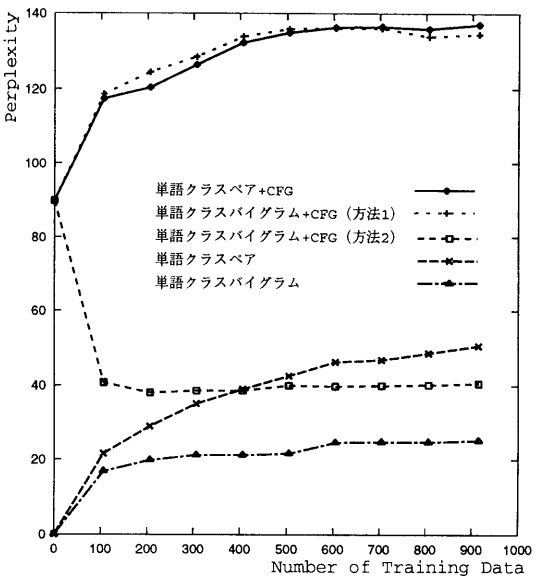

図 6: バイグラムによるパープレキシティの変化
（未登録単語登録後：419 単語）

アと同一である)。また、図6はバイグラムによる効果をグラフにまとめたものである。

単語ペアによる確率の導入の効果は単語バイグラム+CFG の場合（方法 1）は学習セットが少ないためか現段階ではあまり有効とは言えなかった。方法 2 については、バイグラムの予測確率を全ての箇所で使うことによりパープレキシティがかなり改善されることが分かった。また、表 5の場合の様にバイグラムだけで解析を行なうようにすればパープレキシティをかなり改善することができるので、単語ペアの確率化だけでなく CFG 自体の確率化も行なう必要があると言える。

なお、式 (2) のλは本実験では 0.5 としたが、最適値は学習データ文数に依存するので、適応的に更新するのが望ましい。

表 7: バイグラム+CFG での評価結果 (419 単語)
（スムージングなし：括弧内単語クラス）

	方法 1		方法 2	
学習 set	perp.	perp2.	perp.	perp2.
set2	108.2(118.5)	106.3(107.7)	34.9(40.8)	28.5(34.0)
set2-3	114.6(124.4)	106.7(113.6)	30.3(38.0)	24.2(30.8)
set2-4	115.1(128.6)	107.8(120.7)	22.9(38.6)	21.7(31.0)
set2-5	117.4(133.8)	110.9(130.5)	22.7(38.6)	21.7(31.1)
set2-6	118.8(135.8)	116.4(133.7)	23.0(40.0)	20.5(30.5)
set2-7	117.1(136.1)	116.3(133.5)	22.3(39.8)	19.3(28.7)
set2-8	117.1(136.1)	116.3(133.6)	21.9(40.0)	18.8(28.5)
set2-9	117.4(133.8)	116.5(133.2)	21.0(40.2)	17.6(28.4)
set2-10	117.8(134.4)	116.6(133.4)	21.0(40.5)	17.1(28.5)

6 むすび

音声認識システムでユーザが受理できない文を発話した時に、新しい規則を登録し、入力文のカバーレイジを改善するシステムの開発を行ない、" 富士山観光案内システム" のタスクを使うことにより評価を行なった。

CFG の規則の登録法はカバー率ではかなり改善されたがパープレキシティがかなり増加してしまった。非終端記号のペアはトップのペアを使えばカバー率はほぼ完全になるがパープレキシティが大きくなり過ぎ、ボトムのペアを使えばカバー率があまり良くならなかった。一方、単語クラスペアを使った方法はカバー率とパープレキシティのバランスが上記の方法と比べて一番良いということが分かった。

確率を使うことにより単語ペアのパープレキシティの増加を抑えることも試みた。CFG を使わずに確率だけで解析する場合には非常に効果があったが、カバーレイジは十分ではなかった。CFG とバイグラムを組み合わせて使用する方法は、確率を学習するためのデータが不十分な場合にも有効である。CFG で解析できない場合に限りバイグラムを使用する方法は、どんなに効果があったとしても初期の CFG でのパープレキシティよりよくなることはありえず、CFG 自体の確率化 (SCFG) が必要になってくる。一方、バイグラムを CFG と並列に使用する場合はカバーレイジ、パープレキシティの両面において非常に効果があった。

大規模なデータベースを利用できる場合はトライグラム等の確率モデルの方が CFG のような文法よりもパー

プレキシティの面で優れているが、前もって大量のデータが得られないようなアプリケーションではCFGベースの文法の方が利用価値がある。すなわち初期文法はCFGを用い、以後学習データの増加とともにn-グラムへの移行（(2)式のλを$\lambda = 0$から$\lambda = 1$に逐次変更することに相当）というのがアプリケーションによっては現実的であると思われる。本稿で提案した手法は、CFGとバイグラムの両方の利点を活かした文法学習法と言える。

今後、文法規則の学習法で以下のことについて研究を進めていく予定である。

- λの逐次更新法
- CFGのS-CFG化
- 音声認識装置へのインプリメント

現在はテキストレベルでのカバー率を調べている段階なので、実際の音声認識でどれだけ有効かを調べていきたい。

参考文献

[1] 山本 幹雄, 肥田野 勝, 伊藤 敏彦, 甲斐 充彦, 中川 聖一:「自然発話の意味理解と対話システム」 情報処理学会 音声言語情報処理研究会, 94-SLP-2-13, pp.91-98 (1994.7)

[2] 大谷 耕嗣, 山本 幹雄, 中川 聖一:「例文からの話し言葉用文法の半自動修正法」 情報処理学会第50回全国大会, 2R-4, Vol.3, pp.59-60 (1995.3)

[3] 大谷 耕嗣, 中川 聖一:「単語対制約の追加による話し言葉用文法の自動修正法」 情報処理学会第51回全国大会論文集 4H-7, Vol.3, pp.67-68 (1995.9)

[4] 大谷 耕嗣, 中川 聖一:「CFGとバイグラムの結合による文法の半自動修正法」 情報処理学会 音声言語情報処理研究会, 95-SLP-9-15, pp.99-104 (1995.12)

[5] 周旻, 中川聖一:「日本語及び英語の言語モデルに関する検討」, 「自然言語処理における学習」シンポジウム, 電子情報通信学会, pp. 57-64 (1994.11)

[6] 中澤 聡, 濱田 喬:「例文からの学習による生成規則の自動修正」 情報処理学会第49回全国大会論文集 2J-8 (1994.9)

[7] 白井 清昭, 徳永 健伸, 田中 穂積:「コーパスからの文法の自動抽出」 情報処理学会 自然言語処理研究会, 94-NL-101 Vol. 101, pp.81-88 (1994.5)

[8] Eric Brill : Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach, Proc. ACL93, pp.259-265(1993)

[9] Scott Miller, Heidi j. Fox : Automatic Grammar Acquisition, Proc. Human Language Technology, pp.268-271(1994)

[10] G.J.F. Jones, J.H.Wright, E.N. Wrigley: The HMM Interface with Hybrid Grammar-Bigram Language Models for Speech Recognition, Proc. ICSLP-92, pp.253-256(1992)

[11] Christer Samuelsson; Grammar Specialization Through Entropy Thresholds, Proc. ACL94, pp.188-195(1994)

[12] 竹澤 寿幸, 森元 逞:「部分木を単位とする構文規則と前終端記号のバイグラムを利用した連続音声認識」 情報処理学会音声言語情報処理研究会, 95-SLP-9-9, pp.55-62(1995.12)

付録 CFGと単語ペアを使用したトップダウンパーザ

ここでは、第3.2節で報告している方法についての、トップダウンのパーザのアルゴリズムについて説明する。今までは、単語ペアとCFGを並列に解析するパーザは、制作が簡単であったボトムアップによる方法で解析を行なっていた。しかし、音声認識への応用を考えると、我々の音声認識システムがトップダウンで先の単語を予測するパーザを使っている関係から、単語ペアの解析にもトップダウンでのパージングを行なうパーザを開発する必要が生じる。そこでトップダウンのパーザの制作を行なったので、以下そのアルゴリズムについて述べる。

前処理 全ての単語クラスを調べ、各単語毎にその単語を（右辺にもつ）呼ぶ単語クラスのテーブルを作る。また全ての非終端記号をパージングしてみて、どの非終端記号がどの単語クラスを先頭で呼ぶのか調べておく。

入力単語列を$W_1, W_2, \cdots, W_n$とする。

i = 1;

[1]

- i = 1;

  [A] W_iを呼ぶ単語クラスをテーブルから調べて、この単語クラスを先頭で予測する生成規則の解析の途中結果を覚えておく。

- i != 1;

  [B] 全てのパージング途中の結果を一単語進めてみる。その時、パージングが終了したものがあり単語ペア(W_{i-1}, W_i)があれば[A]の操作を行なう。パージングが成功していないものについては、予測単語とW_iとのチェックを行ない一致した候補については残し、一致しないパージング結果は捨てる。

[2]

- i = n;
  候補のパージング途中の結果のうち、一単語解析を進めた時に、解析が成功している候補があれば解析成功、なければ失敗とする。

- i != n;
  i = i + 1, [1]へ。