音声言語情報処理　9－15
(1995. 12. 15)

CFG とバイグラムの結合による文法の半自動修正法

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) 47-0111
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半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をし、その結果文のカバー率が改善される。

未登録単語の登録法以外に文法の学習に2つの方法を提案する。1つは、文脈自由文法の規則の追加学習法で、もう一つは単語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と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, カバー率, パープレキシティ

A Semi-Automatic Learning Method of Grammar Rules by the Combination of CFG and Bigram

Kouji OHTANI and Seiichi NAKAGAWA

Department of Information and Computer Sciences
Toyohashi University of Technology
Tenpaku-cho, Toyohashi, 441, Japan
Tel. (0532) 47-0111
Email {otani, nakagawa}@slp.tutics.tut.ac.jp

[abstract] In this paper, we describe a semi-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 two methods for learning the grammars. One is to add new production rules to the original CFG. The other is to add word pairs. After learning unknown words and production rules, we evaluate the coverage of sentences and perplexity using above methods.

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

1　序論

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法として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]。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 [4]、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 [5]。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、その文法の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 [6]。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 [7]。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な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 [8]。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 [9]。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 [10]。

本システムは文法を学習するために 2 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。もう一つは接続可能な単語ペアを登録する方法である。未登録単語の登録ではトップダウンのパーザを使って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）であるワードクラスの中から適当なものを見つけ、その単語クラスの中に未登録単語を登録する。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生成規則または単語ペアを登録する方法である。

未登録単語の登録について、”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、CFG の方法では評価の結果、受理可能な文の数は向上したがパープレキイティがかなり増加してしまった。一方、単語ペアの登録方法では、受理可能な文の数の増加は CFG の場合とほぼ同じであるが、パープレキシティの増加が抑えられた。

2 節でシステムの認識エラーの原因について述べ、3 節で未登録単語と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述べる。4 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）の時のエラーである（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（もしくは単語クラスのペア）

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 番目のエラーは、現在のところ扱っていない。

3 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。文法の学習には 2 つの方法を考察した。1 つは未登録単語のための単語クラスへの登録及び未登録単語自身の登録で、もう 1 つは規則の欠落のための生成規則の登録である。未登録単語の登録はトップダウンのパーザを使い未登録単語のための適当な単語クラスを見つけ、その単語クラスへ未登録単語の登録を行なう。生成規則の登録は文法の学習に 2 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、もう 1 つ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ある。

3.1 未登録単語の登録

未登録単語の登録手順を以下に述べる。

1. 未登録単語が入力された時、仮に任意の単語クラスにその未登録単語が含まれいると仮定し、一つづつ順に所属する単語クラスを仮定する。

2. 入力文の解析が成功する単語クラスの仮定が 2 つ以上ある場合は、ユーザがふさわしい候補を選ぶ。

図 1は未登録単語を登録するアルゴリズムを示している。入力文が 1 つの未登録単語のために受理できない時、システムは与えられる部分単語ストリングのための次の単語又は単語クラスを予測できるトップダウンのパーザを使う。

以下に単語登録の詳細を述べる。解析で入力文中にただ 1 つの未登録単語が含まれている時、その単語が任意の単語クラスに属するとする。そして入力文の解析を始め、解析が未登録単語の所まできた時、その時に予測されている単語クラスが未登録単語の単語クラスの候補であると仮定する。そしてシステムは仮定された単語クラスを使って解析を続け、入力文の解析が終了した時に解析が成功したとする。システムは解析に成功した仮定された単語クラスの候補とその単語クラス内の前

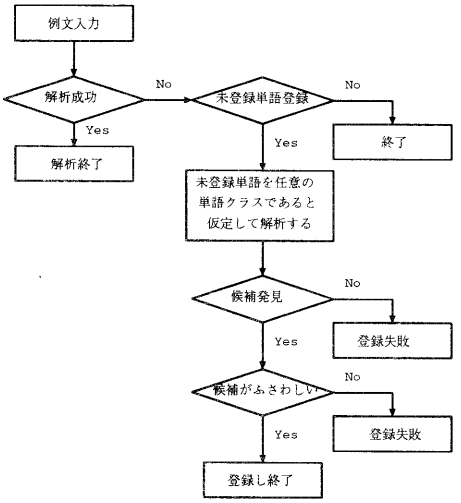

図 1: 未登録単語の登録アルゴリズム

もって登録されている単語を使うことによっていくつかの例文を生成し、これらの例文の妥当性を判断することにより解析に成功した単語クラスの候補が適当かどうかをユーザに判断してもらう。もし候補が適当ならば文法中に登録される。以上のアルゴリズムを使って未登録単語の登録が実行される。

3.2 生成規則の登録

生成規則の登録には 2 つの方法がある。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、もう 1 つは部分的にパーズされた各ストリングの間の単語ペア又は単語クラスのペアを使う方法である。2 つの方法について以下に述べる。

3.2.1 CFG での規則生成

CFG の生成規則を登録する時、システムは部分解析木を作るボトムアップパーザを使い、図 2で示されるような以下のステップで登録が行なわれる。

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以下で述べられるような図 3 に示す 3 つの登録方法を使い分ける。
4. 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る。

図 3に図 2で使われている 3 つの登録方法の例を示す。

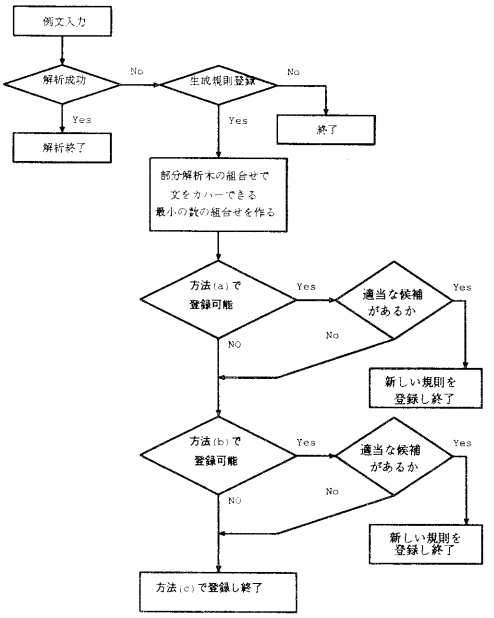

図 2: 生成規則の登録アルゴリズム

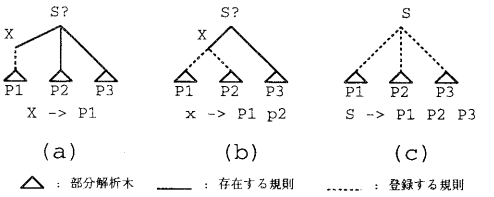

図 3: 生成規則登録の際の 3 つの方法

(a) ある部分解析木の非終端記号を適当に変えた時、入力文が受理できるかどうかを調べる。例えば、図 3 (a) で $X \rightarrow P1$ を登録した時、新しい規則の登録が修正されたシステムの文法を使ってこの入力文が受理可能かどうかを調べる。

(b) 2 つの部分解析木の非終端記号をつなげた時、入力文が受理できるかどうかを調べる。例えば、図 3 (b) で $X \rightarrow P1\ P2$ を登録した時、新しい規則の登録によって修正されたシステムの文法を使ってこの入力文が受理可能かどうかを調べる。

(c) 部分解析木の最小の組合せを直接開始記号につなげる。例えば、図 3 (c) で開始記号に直接の接続、例えば $S \rightarrow P_1P_2P_3$ を作りそれを登録する。

登録の優先順位は $(a) > (b) > (c)$ である。なぜならこの順序が特殊化のしすぎを小さくすると思われるか

らである。

3.2.2 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。単語又は単語クラスペアを使う登録には3つの方法がある。1つ目は、CFG規則の補助として単語対制約を登録する方法（方法1）で、以下の手順で登録を行なう。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。

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4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。

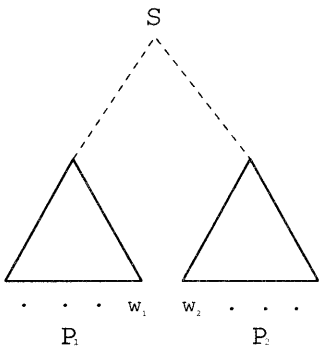

図4: 単語対制約の登録

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。

もう1つの方法（方法2）は、登録方法については方法1とほとんど同じで入力文に対して最小の数でカバーできる部分解析木の組合せを見つけた上で、この組合せについて部分解析木に隣接する単語（単語クラス）のペアを登録するとともに、この部分解析木の組合せを開始記号で書き換える規則も登録する（図3の（c）に対応）。つまり、図4の場合だと単語ペア (w_1, w_2) とCFG規則 $S \rightarrow P_1 P_2$ の両方を登録する。ただし、この新たに登録するCFG規則は、その適用時に非終端記号間で単語（単語クラス）ペアのチェックを行なう制約付きの規則とする。方法1と方法2の違いは、方法1のペアは任意の規則間で適用が可能なのに対して、方法2では新たに登録した制限付きCFG規則の規則でしか使用できない点である。

また、比較のため方法3として入力文から接続可能な単語のペアを獲得して、CFGを使わずに獲得した単語ペアだけで文の解析をする方法も行なった。

4 評価

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。

- 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
- 前終端記号から終端記号への書き換え規則数- 255
- パープレキシティ - 76.5

学習と評価で使われたデータは過去の研究[11]集めた335文と今回新たに集めたデータ685文の計1020文を使用した。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、今回の学習では対象外とした文の数を示している。表中のsetのうちset1を評価にその他のsetを学習に使った。これより大まかに言えば、受理可能になる文数は39文＋35文＋α＝74文＋αが本研究の上限である（αは未登録単語が登録されることによって増える）。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）／（set 1のうち未登録単語の含まれていない文の数）を示している。”単語”は（学習後のset 1の未登録単語の異なり数）／（未登録単語出現数）を示している。表3と表4についてもこれらの4つのうち解析できた文の数以外は同じなので2つの表では省略している。”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。表3と表4は単語（括弧内は単語クラス）ペアについて方法1と方法2の学習の結果を示している。表中の’ペア数”は学習・登録された単語（単語クラス）のペアの数を示している。登録規則中の”＋1”は、最小の部分解析木の組合せが1となったもので、これは通常のCFGとして登録したためこのような表記となっている。表5は単語ペア（方法3）の学習の結果を示している。表中の”

単語数" は出現単語の異なり種類の数を示している。括弧内は表2での" 単語" 欄に対応する。また" カバー数" も表2の" カバー数" 欄に対応している。表3のパープレキシティについては、オリジナルな元の初期CFGで受理可能だった学習セット中の39文を使用して求めた。これは、この方法についてはトップダウンでの解析パーザができていないためである。

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。

図3に対応する3つの登録方法のうちで、方法(c)によって実行されたCFGでの生成規則の登録が約70%あった。文のカバーレイジが増加するに従ってパープレキシティも増加してしまった。パープレキシティはオリジナルな初期文法で解析可能な50文を使い、各登録方法毎に求めている。パープレキシティの増加に関しては方法(c)による登録が最も影響が大きい。これは登録規則の数が一番大きいから当然でもある。しかし、方法(a)による登録数は少ないにもかかわらず、パープレキシティはかなりの増加があった。方法(b)による登録は登録数が少ないのでパープレキシティの増加も少なかった。

単語（単語クラス）のペアによる学習の効果は方法1についてはCFGを使った登録よりも良いと思われる。この方法では最終的に単語クラスを使った場合だと最終的に評価の対象となるセット1の74文中62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。ただし、この方法はボトムアップでの解析をしており本研究室の連続音声認識システムに適用するためにはトップダウンでの解析の方法を考えなければならないのが問題である（このためパープレキシティを39文を使って求めている）。方法2についてはカバー数でもパープレキシティでも方法1よりも悪い結果となってしまった。パープレキシティが方法1よりも悪い原因として、適用条件は方法1よりも厳しいが、規則の追加が増えたためだと思われる。方法3は単語ペアだけでの学習法で、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。

単語クラスのペアを使った方法の方が単語ペアの方法よりも当然カバーレイジが優れている。他方で、パープレキシティでは当然単語クラスペアの方法の方が単語ペアの方法よりも大きくなる。しかし、パープレキシティの増加は10%も多くはない。単語クラスペアの登録による方法は方法1では学習前の1.2倍強程度でであることから、単語クラスのペアの登録法は有効であると思われる。全ての方法の中で、トップダウンでの解析が可能になれば方法1の単語クラスのペアを使う方法が最も良い結果であるといえる（現在解析法を検討中である）。また、学習セットを増やしていけば方法3の単語ペアだけの方法、または確率を考えたバイグラム等が有効になると思われる。

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

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	単語	perp.	規則数
set2	52/74	33/47	111.2	30
set2-3	54/74	33/47	119.0	45
set2-4	58/74	33/47	130.8	65
set2-5	58/77	32/43	136.5	78
set2-6	58/77	32/43	138.2	98
set2-7	-/77	32/43	—.-	114
set2-8	-/77	32/43	—.-	135
set2-9	-/77	32/43	—.-	154
set2-10	-/78	31/41	—.-	170

表3: 単語ペア＋CFG（方法1）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.	ペア数
set2	45(47)	78.0(78.6)	60(67)
set2-3	50(54)	78.7(79.6)	83(94)
set2-4	51(57)	78.9(82.1)	103(127)
set2-5	52(58)	79.2(83.4)	123(149)
set2-6	53(59)	79.5(84.9)	144(170)
set2-7	54(61)	79.7(85.4)	164(186)
set2-8	54(61)	79.9(85.7)	183(200)
set2-9	54(61)	80.2(86.3)	206(228)
set2-10	54(62)	80.2(86.3)	216(233)

表4: 単語ペア＋CFG（方法2）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.	ペア数	規則数
set2	42(43)	83.9(82.3)	60(67)	28
set2-3	47(47)	95.4(93.7)	83(94)	44
set2-4	47(50)	95.1(97.2)	103(127)	55
set2-5	47(50)	96.9(100.5)	123(149)	69
set2-6	47(50)	97.7(101.2)	144(170)	88
set2-7	49(51)	100.5(104.4)	164(186)	102+1
set2-8	49(53)	103.2(110.3)	183(200)	118+1
set2-9	49(53)	104.6(113.0)	206(228)	133+1
set2-10	49(54)	106.7(113.0)	216(233)	141+1

表 5: 単語ペア（方法 3）の学習結果

	単語数	カバー数	perp.	ペア数
set2	160(51/96)	12/43	6.9	391
set2-3	203(32/38)	21/77	8.9	598
set2-4	256(19/21)	30/88	10.5	837
set2-5	288(17/18)	39/91	12.8	1024
set2-6	306(17/18)	42/91	13.8	1152
set2-7	320(17/18)	45/91	14.5	1292
set2-8	329(16/17)	48/92	14.9	1366
set2-9	347(15/16)	51/93	16.5	1491
set2-10	359(13/13)	52/96	17.3	1583

5 結論

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った。CFG を使った生成規則の登録では、適当な一般性を与えるために 3 つの方法を使い分けるようにした。また、単語又は単語クラスのペアの登録についても考察した。評価は" 富士山観光案内システム" のタスクを使うことにより行なった。未登録単語の登録については約 1000 文中 46 文（22 単語）の登録を行なったが、現在の方法は未登録単語が文中に 1 つの時にしか使えない上、生成規則もそろっていなければ登録することができず完全とはいい難い。

単語又は単語クラスのペアの登録については、CFG の規則の登録法と比べてカバー率では若干劣るもののパープレキシティでの増加は少なく、全体として改善がみられた。単語ペアの登録・使用法について 3 つの方法を調べたが、単語クラスのペアを使った方法 1 での学習が一番良い結果を与えるといえる。

大規模なデータベースを利用できる場合は trigram 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。本稿で提案した手法は、この両方の利点を活かした文法学習法と言える。

また未登録単語と CFG の規則の登録には、適当な候補を選択するためにユーザの判断が必要であるから、文法の獲得は半自動的に行なっている。将来的には文法規則の登録で以下の点を改善していきたい。

- カバー率の改善
- パープレキシティの削減
- 自動登録化

これらの問題を解決するために、大きなデータコーパスから確率付の文法を自動的に学習することについて考えていきたい。また今回文法の学習の対象外となった生成規則の不備と未登録単語が交じった文や 1 文中に複数の未登録単語が存在する文についての学習についても検討したい。

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] 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)

[5] 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)

[6] 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)

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

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

[9] 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)

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

[11] 伊藤・大谷・肥田野・山本・中川：「事前説明によるシステムへの入力発話の変化と認識結果の人間による復元」 情報処理学会音声言語情報処理研究会, 94-SLP-4-7, pp.49-56(1994.12)