4H－7

単語対制約の追加による話し言葉用文法の自動修正法

大谷耕嗣　中川聖一
豊橋技術科学大学 情報工学系

1　はじめに

これまで、我々の研究室では音声による対話理解の研究として富士山の観光案内をタスクとするシステムの開発を行なってきた。このシステムはユーザが話した文を認識するためにユーザが話すであろう文を文脈自由文法(CFG)で記述している。しかし、システムの文法で受理されない文をユーザが発話した時には、当然システムは、ユーザが話した言葉を認識・理解することはできない。これが、システムの認識間違いの原因の一つとなっている。

そこで、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登録を行ない、これらの文の理解を可能にするシステムの開発を行なってきた[1]。この登録方法により、解析できなかった文が解析できるようになり、カバー率は増えたが、パープレキシティ（2のエントロピー乗：ある部分文に接続可能な単語数）がかなり増加してしまった。今回は、システムのカバー率ををあまり減らさずにパープレキシティの増加を押えることを、CFGに単語対制約を新たに加えることにより実現することを試みた。

2　文法の登録

ユーザがシステムに入力した文の単語が、システムの文法に未登録であるために解析できない文のための未登録単語の登録は文献[1]の方法を用いて行なう。また、入力文の単語はすべてシステムの文法に登録されているが生成規則が未登録であるために解析できない文については、CFG規則の登録を行なった場合[1]と単語対制約の登録を行なった場合の比較を行なった。単語対制約を使う登録は2通りの方法を行なった。単語対制約だけをCFG規則の補助として使用する方法と、新たなCFG規則とそれに伴う単語対制約を登録する方法である。また、単語対制約については単語のペアを考えた場合と単語クラスのペアを考えた場合それぞれについて比較を行なった。

一つ目は、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。

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その入力文をカバーす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。

2. 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。

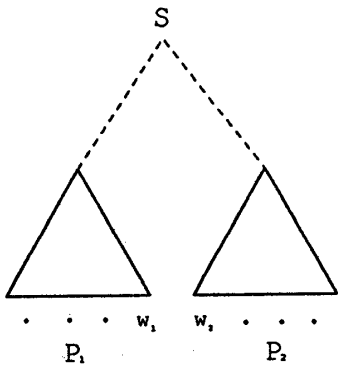

図1: 単語対制約の登録

この方法で登録した単語対制約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と単語ペアを独立に用いて後続単語の予測を行なう。

もう一つの方法（方法2）は、登録方法については1つ目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。つまり、図1の場合だと単語対制約(w_1, w_2)とCFG規則$S \rightarrow P_1P_2$の両方を登録する。ただし、この新たに登録するCFG規則は非終端記号間で単語対制約のチェックを行なう制約付きの規則とする。即ち、方法1とは違い単語対制約は新たに登録したCFG規則の適用時のみ制約がかかる。

An Automatic Correction Method of Grammar Rules for Spontaneous Speech by Word Pair
Kouji Ohtani, Seiichi Nakagawa
Toyohashi University of Technology

3 評価

実験には、本研究室で作成している富士山の観光案内をタスクとする対話システムで使用している文法を用いた。文法の細かい内訳は単語数241、生成規則数391、非終端記号135、単語クラス数110、単語クラスから単語への書き換え規則数255、パープレキシティ75.9となっている。

学習と評価に使ったデータは文献[2]で集められた計10人の話者のデータ（セット1、2、各106文）と追加実験で集めた10人の話者のデータ（セット3、123文）からなる。

表1は学習セットの詳細を示している。この表で示されている解析失敗の原因が単語の未登録である文、および生成規則が未登録である文について規則の登録を行なった（表1の単語&生成規則の欄に相当する文は対象外)。表2は登録した単語および単語クラスペアの数を示している。表3は、文法を登録することによって解析可能となった文の数を示している。表4は、文法を登録する前と登録した後のパープレキシティの変化を示している。また、表3、4での学習後の欄のCFGは、文法をCFGのみで学習した場合[1]の結果で、単語対制約の欄は、単語ペアと単語クラスのペアで学習した結果についてまとめている。( )内の値が単語クラスのペアを学習した結果である。

表1: 学習セットの詳細

	受理文	解析失敗の原因		
		単語未登録	生成規則未登録	単語 & 生成規則
set 1	39	0	35	32
set 2	51	5	32	18
set 3	70	4	28	21

表2: 登録した単語ペアの数

	set1	set2	set3	set1+3	set2+3
word pair	65	60	58	108	98
wordclass pair	59	58	61	103	95

表3、4より、方法1についてはCFG規則の登録より文カバー率が若干の減少でパープレキシティをかなり減らすことができた。しかし、方法2は方法1に比べてカバー率でもパープレキイティでも悪い結果となってしまった。これは、規則間の制約は強くしたがCFG規則を登録したために文の最初に対する予測される単語の数が増えてしまったためにパープレキシティが増えてしまったためと考えられる。また、単語対制約を登録する方法は学習データが少ないために、まだカバー率が悪いと考えられるので学習データを増やす必要があると思われる。

表3: 文カバー率の改善結果

学習セット	評価セット	受理可能文			
		学習前	学習後		
			CFG	単語対制約	
				方法1	方法2
set 1	set 2	51	60	57(57)	55(55)
set 3	set 2	51	63	61(62)	58(59)
set 1+3	set 2	51	68	63(65)	59(61)
set 2	set 1	39	52	45(47)	43(43)
set 3	set 1	39	50	46(49)	44(44)
set 2+3	set 1	39	56	51(55)	47(47)

表4: パープレキシティの変化

学習セット	学習前	学習後		
		CFG	単語対制約	
			方法1	方法2
set 1	75.9	90.7	76.6(77.6)	78.3(80.4)
set 1+3	75.9	109.0	77.1(78.9)	83.2(87.1)
set 2	75.9	106.6	78.9(79.5)	80.2(82.6)
set 2+3	75.9	123.3	80.4(81.6)	83.4(87.2)
set 3	75.9	101.8	77.7(78.7)	79.9(81.1)

単語対制約を使った場合での、単語ペアと単語クラスペアとの違いはあまりなかった。単語ペアの方がカバー率が若干下がる代わりにパープレキシティでは多少よい結果が得られた。結局、すべての方法の中で方法1の単語対制約を使い単語クラスのペアを学習する方法が一番良い結果が得られた。

4 むすび

対話システムにユーザが受理できない文を話した時に、その文の登録されていない規則を登録することにより新たな文の解析を可能にするシステムの開発を行なった。従来のCFGでの学習に単語対制約を加えることにより文カバー率およびパープレキシティの改善を調べた。その結果、カバー率の多少の減少で、パープレキシティにかなりの改善がみられた。現在、さらにシステムの改善を目指し、今回登録の対象外であった、解析できなかった原因が単語の未登録と生成規則の欠落にあった文の登録について検討している。

参考文献

[1] 大谷, 山本, 中川:「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.59-61 (1995.3)

[2] 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 (1994.12)