TR-I-0141

テキスト･データベースを用いた 文脈自由文法の適用確率推定
Estimation of Grammar Probabilities from Text Database

重盛 勝† 北 研二 森元 逞
Masaru Shigemori Kenji Kita Tsuyoshi Morimoto

1990.3

概要

テキスト･データベースを用いて、日本語の文節構造を記述した文脈自由文法の適用確率を推定した。推定方法としては、構文解析木の出現確率をもとに推定する方法を用いた。この文法から自動的に生成した確率LR構文解析表をHMM-LR音声認識システムに組み込んで音声認識実験を行った結果、第1位での認識率が、平均で、以前の確率文法(規則適用回数の頻度によって確率を推定していたもの)を用いた場合の92.1％から、92.9％へ向上した。

ATR自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

† 豊橋技術科学大学 情報工学課程

目次

1. はじめに

ATR自動翻訳電話研究所では、自動翻訳電話の実現に向けてHMM-LR音声認識システムを開発してきた[1][2]。このシステムの音声認識精度を向上させるために、統計的な言語情報を用いる研究が行われている。その1つに、SHIFT/REDUCEに確率の与えられたLR構文解析表(以下、確率LR構文解析表と呼ぶ)を用いる方法があり、これによる認識精度の向上が報告されている[3]。

この方法では、確率LR構文解析表の確率に、言語情報がより正確に反映されているほど高精度の音声認識を行うことができる。確率LR構文解析表は、確率文脈自由文法から自動的に生成されるので、確率文脈自由文法の確率を、より正確に言語情報を反映するように求めることが要求される。

確率文脈自由文法の確率は、あらかじめトレーニングデータを使って推定することができる。確率文脈自由文法の確率推定方法としては、Forward / Backwardアルゴリズムを応用したOutside / Insideアルゴリズムが知られている[7]。これは、すべての可能な導出を考慮しながら確率を推定する方法で、1つのデータに対して数多くの解釈ができる場合に有効である。一方、ATRのHMM-LR音声認識システムで認識対象としている日本語の文節構造のあいまいさは小さい。そこで、O/Iアルゴリズムを使わなくても、データを効率よく構文解析して、その結果を利用して確率を推定することができる。

従来の確率LR構文解析表の確率は、トレーニングデータを構文解析するときに文法規則が使われた回数の頻度として求められていた(今後、本レポートで以前の確率とある場合はこの確率をさすものとする)。しかし、あいまいなデータを解析したときの解析木の出現確率をもとに確率を推定する方法が、Fujisakiら[4][5]によって紹介されており、この方法によって推定した確率の方がより正確に言語情報を反映する。

本レポートでは、日本語の文節構造を記述した文脈自由文法の確率を、構文解析木の出現確率をもとに推定した結果について報告する。また、その文脈自由文法から生成した確率LR構文解析表を用いて行った音声認識実験の結果について報告する。

日本語のテキスト・データベースを用いて確率推定した文法の複雑度は、以前の3.12から3.10へ改善された。また、推定した確率を用いた音声認識実験では、日本語の文節単位の音声に対する第1位の認識率が平均して以前の92.1%から92.9%へ向上した。

2. HMM-LR音声認識システムにおける確率LR構文解析表の利用

高精度の音声認識を行うためには、言語情報の利用が不可欠であり、そのためにATRではHMM-LR法を用いた音声認識システムを開発してきた[1][2]。このシステムの音声認識精度を向上させるために確率LR構文解析表 を用いる方法がある。

本章では、まずHMM-LR音声認識システムについて簡単に説明する。次に、このシステムに、確率LR構文解析表を用いる方法を説明する。

2.1 HMM-LR音声認識システム

図2.1に、HMM-LR音声認識システムを示す。システムは大きく分けて、HMM音韻照合部(HMM Phone Verifiers)と予測LRパーザ(Predictive LR Parser)から構成される。

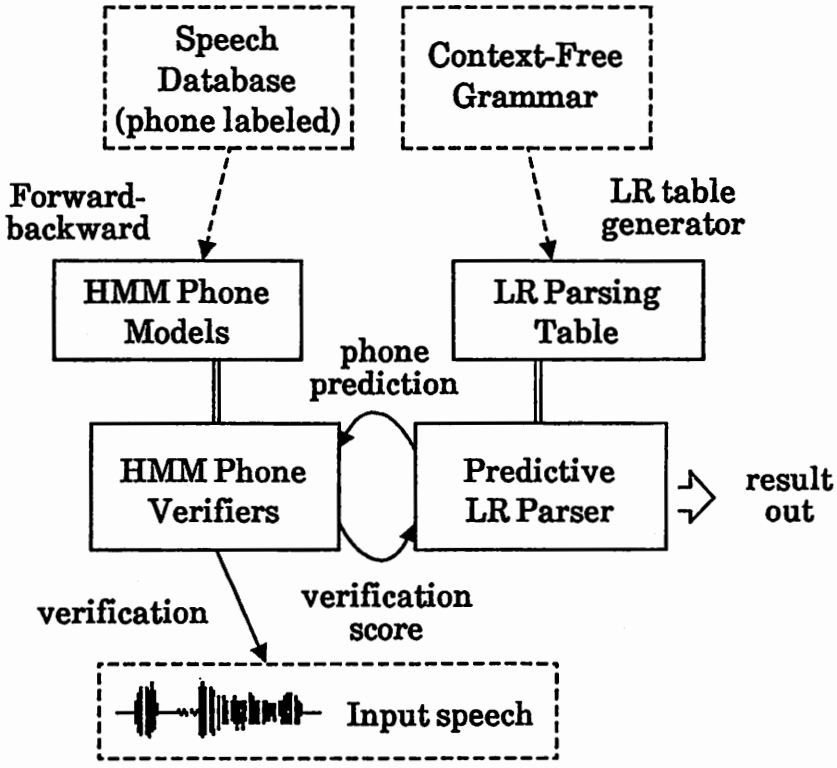

図2.1 HMM-LR音声認識システム

HMM音韻照合部は、あらかじめ学習された音韻のモデルを用いて音韻照合を行い、指定された音声区間に対する音韻の照合スコアを計算する。予測LRパーザは、LR構文解析表を参照しながら、文法上で次に接続しうる音韻を予測する。

HMM-LR法では、次のようにして音声認識を行う。

(1) 予測LRパーザがLRテーブルを参照して文法上で次に接続しうる音韻を予測し、それらの音韻に対してHMM音韻照合部を駆動して照合スコアを求める.

(2) 照合に成功したすべての音韻に対して、並行して音韻連鎖の枝をのばしていき、これらの枝の間でビームサーチを行いながら認識を進める。

通常のLRパーザでは、入力記号列を与えてLRテーブルを用いて解析するが、HMM-LR法では逆に予測LRパーザがLRテーブルによって次に来る音韻の予測を行いながら音声認識を行う。

HMM-LR法では、音韻ラティスなどの中間形式を介さずに認識が進むので、照合スコアの計算が正確になり、高い認識性能が得られる。

2.2 HMM-LR音声認識システムにおける確率LR構文解析表の利用 [3]

HMM-LR音声認識システムは、LR構文解析表を用いて文法的な制約によってサーチ・スペースを縮小している。従来のLR構文解析表では、LR構文解析表の各状態においてSHIFTとREDUCEが等しく扱われていた。しかしながら、ある動作は頻繁に起こるが他の動作はめったに起こらないというような現象があると考えられる。この現象を扱うために、LR構文解析表のSHIFT/REDUCE動作に確率の与えられた確率LR構文解析表を使うことが考えられた。確率LR構文解析表の例を図2.2に示す。

state	o	e	u	k	r	$	S	NP	N	V	P
0	<s3,0.18>			<s4,0.82>			1	5	6	2	
1						<acc,1.0>					
2						<r2,1.0>					
3				<s7,1.0>							
4	<s8,0.85>		<s9,0.15>								
5	<s3,0.6>			<s11,0.4>						10	
6	<s13,0.8>			<r3,0.2>							12
	<r3,0.2>										
7			<s14,1.0>								
8					<s15,1.0>						
9					<s16,1.0>						
10						<r1,1.0>					
11			<s9,1.0>								
12	<r4,1.0>			<r4,1.0>							
13	<r6,1.0>			<r6,1.0>							
14					<s17,1.0>						
15		<s18,1.0>									
16		<s19,1.0>									
17		<s20,1.0>									
18	<r5,1.0>			<r5,1.0>							
19						<r7,1.0>					
20						<r8,1.0>					

図2.2 確率LR構文解析表の例

この確率LR構文解析表を用いることにより、HMM-LR音声認識システムにおいて認識候補である各音韻連鎖の確率は、その時点までに行われたLRの動作に与えられた確率の積として求めることができる。そして、この確率の低い候補を除くことによってサーチ・スペースをさらに縮小できる。これは、探索候補の絞り込みに言語的確からしさを加えていることになり、認識精度が向上することが期待できる。

この方法では、確率LR構文解析表の確率に、言語情報がより正確に反映されているほど高精度の音声認識を行うことができる。確率LR構文解析表は、確率文脈自由文法(これについては3章で詳しく述べる)から自動的に生成されるので、確率文脈自由文法の確率を、より正確に言語情報を反映するように求めることが要求される。

3. 確率文脈自由文法の定義

確率文脈自由文法は図3.1に示すように4つ組で定義される。

確率文脈自由文法 $\equiv (V_N, V_T, Ps, S)$

V_N … 非終端記号の集合,
V_T … 終端記号の集合,
Ps … 書き換え規則の集合,
S … 開始記号,

$$Ps = \{ < \alpha \rightarrow \beta, p > | \alpha \in V_N, \beta \in (V_N \cup V_T)^*, \quad p = P(\beta|\alpha), \Sigma_\beta P(\beta|\alpha) = 1 \}$$

図3.1 確率文脈自由文法の定義

各書き換え規則には確率が与えられている。これは、ある文法記号がその文法規則によって別の文法記号列に書き換えられる確率である。左辺が同じ文法規則の確率を足し合わせると1になる。図3.2に確率文脈自由文法の例を示す。

確率文脈自由文法では、ある文が導出される確率は、その文の導出に用いられた文法規則に与えられている確率の積となる(図3.3参照)。これは、個々の規則適用が文脈に独立に行われるためである。

(1)	S → NP V	0.7
(2)	S → V	0.3
(3)	NP → N	0.2
(4)	NP → N P	0.8
(5)	N → k o r e	1.0
(6)	P → o	1.0
(7)	V → k u r e	0.4
(8)	V → o k u r e	0.6

図3.2 確率文脈自由文法の例

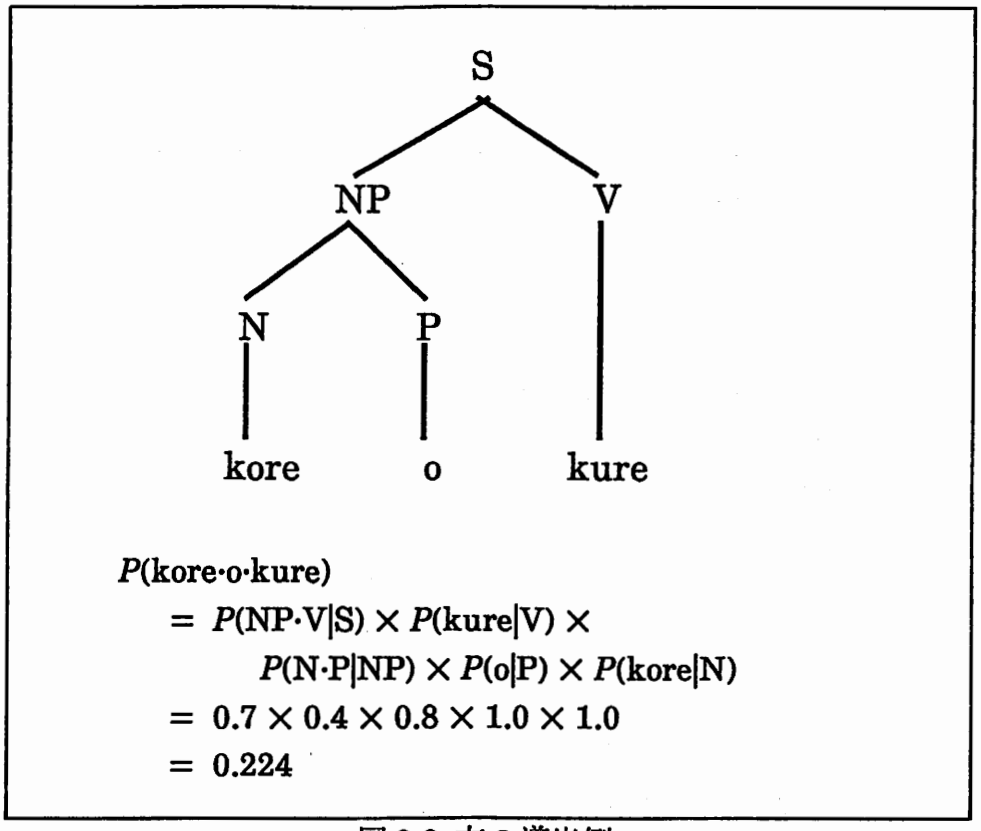

図 3.3 文の導出例

4. 解析木の出現確率をもとにした確率文脈自由文法の確率推定方法

本章では、確率文脈自由文法の確率を、トレーニングデータを使って推定する方法のうち、解析木の出現確率をもとにして推定する方法を説明する [4] [5]。

前提として文法はあいまいであるものとする。したがって、1つの文に対して1つ以上の解釈ができるものとする。

記号の約束として非終端記号αから文形式(非終端記号や終端記号の列)βへの書き換え規則をα→βと表し、規則α→βの適用確率を $Pr(\beta|\alpha)$ と表すことにする。また i 番目のトレーニングデータ B^i を構文解析したときにできる解析木のうち j 番目のものを D^i_j と表すことにする。

$Pr(\beta|\alpha)$ は、以下の繰り返し処理によって自動的に推定することができる。

Step 1: $Pr(\beta|\alpha)$ の初期設定をする。このとき同じ左辺をもつ規則の確率の合計が1になるように次式を満足させる。

$$\sum_{\beta} Pr(\beta|\alpha) = 1$$

Step 2: データ B^i を構文解析する。

Step 3: 解析木 D^i_j の出現確率 $Pr(D^i_j)$ を解析木 D^i_j を得るのに使われた規則 $\{r\}$ の確率の積として計算する。すなわち、

$$Pr(D^i_j) = \prod_{r \in D^i_j} Pr(r)$$

Step 4: データB^iを発生するのに規則α→βが使われた回数の後天的推定(Bayes *a posteriori* Estimate) $C_\alpha{}^i(\beta)$を計算する。

$$C^i_\alpha(\beta) = \sum_j \{ \frac{Pr(D^i_j)}{\sum_k Pr(D^i_k)} \times n^i_j(\alpha,\beta) \}$$

ここで、$n^i_j(\alpha,\beta)$は、規則α→βがD^i_jを発生するのに実際に使われた回数である。

Step 5: 確率の正規化を行う。すなわち、同じ左辺を持つ規則の確率の合計が1になるようにする。

$$f_\alpha(\beta) = \frac{\sum_i C^i_\alpha(\beta)}{\sum_i \sum_\gamma C^i_\alpha(\gamma)}$$

Step 6: $f_\alpha(\beta)$で$Pr(\beta|\alpha)$を更新し、Step 3から繰り返す。

この繰り返し過程によって$Pr(\beta|\alpha)$は収束していくことが証明されている[6]。

このアルゴリズムによって推定された確率が言語的確からさを反映する理由について述べる。

Step 4において、式中の右辺第一項は、解析木の出現における相対確率を表している。すなわち、文があいまいであるときに、その文を構文解析して得られた複数の解析木の中で、1つの解析木が現れる確率を表している。これはあいまいな解釈における信頼度を表すと考えることができる。この信頼度に、実際に規則が使われた回数を乗じ、各解析木の和をとって規則が使われた回数の後天的推定$C^i_\alpha(\beta)$としている。この過程において、

- (文字通り)比較的よく解釈されそうな仕方でできた解析木に使われた回数は重視され、
- 逆にめったに解釈されそうにない仕方でできた解析木に使われた回数は軽視される

ことになる。これによって、推定された確率に言語的確からさが反映されるようになる。

5. 実験

5.1 実験の意図

確率推定方法を改良して文脈自由文法の確率を推定し、それから生成した確率LR構文解析表をHMM-LR音声認識システムに組み込むことにより認識精度が向上するかを調べる。

5.2 実験手順および実験条件

ここでは、実験手順および実験条件について述べるが、実験条件の詳細については、文献[2][3]を参照のこと。

(a)文脈自由文法の確率推定

文法をプログラムに読み込み、トレーニングデータを通して確率を推定する。

文法

以前からHMM-LR音声認識システムにおいて使用されてきたものを使う。これは、日本語の文節構造を規定するもので、きわめて広範囲の言語現象を取り扱っている。表5.1に、この文法の大きさを示す。

表5.1 使用する文法の大きさ

規則数	1,461
語彙規則数	1,320
異なり語数	1,035
LRステート数	4,359

トレーニングデータ

約73,000の文節データ(約300,000音節)から成る日本語のテキスト・データベースを利用する。

プログラム

4節で述べた方法をプログラミングしたルーチンを、以前の確率推定に使用したプログラムに組み込んで改良したものを用いる。処理の大まかな流れは、文法を読み込んで、トレーニングデータを構文解析し、その結果から確率を推定するといったものである。プログラムは、すべてC言語で記述してある。

(b)確率推定した文脈自由文法を用いた音声認識実験

(a)で確率を推定した文法からLR構文解析表を生成し、それをHMM-LR音声認識システムに組み込んで、日本語の文節単位の音声を認識する実験を行う。

音声データ

特定の話者(男性3名、女性1名)によって文節単位に発声された25文章279文節の音声をA-D変換したものである。なお、話者それぞれのデータを区別するために、話者のイニシャルを用いることにする(男性A.U, H.T, N.M, 女性S.U)。

LR構文解析表

(a)で確率を推定した文脈自由文法からプログラムによって自動的に生成する。

5.3 実験結果および結果の検討

(a)文脈自由文法の確率推定

今回確率推定した文法と以前の確率文法を比較するために複雑度(Perplexity)を計算した。複雑度は、一般的な言語モデルとしての良さを表すもので小さいほど言語モデルとして優れていることになる[8]。それぞれの文法の複雑度を表5.2に示す。

これより、確率推定方法を改良した文法の方が複雑度がわずかながら小さくなっており、言語モデルとして良くなっていることがわかる。

表5.2 文法の複雑度

以前の確率文法	改良した確率文法
3.12	3.10

(b)確率推定した文脈自由文法を用いた音声認識実験

実験の結果を音声認識率によって表す。HMM-LRシステムでは候補として残った音韻列に対して、音韻の照合スコアや確率により順位が付けられる。ある順位における音声認識率とは、正しく認識された音韻が、その順位までの候補に入っている割合のことである。

音声認識実験の結果を表5.3～5.7に示す。比較のため、表には、以前に行われた同じデータに対する音声認識実験の結果も入れてある。表5.3～5.6の中で、カッコ内の数値は、正しく認識された音韻がどの順位に入っているかを表している。

これらより、HMM-LRのみの場合よりも、確率文法を用いた場合の方が認識精度が良くなることがわかる。また、改良した確率文法による結果の方が、以前の確率文法による結果よりも少し良くなっている。

個々のデータについて具体的にみると、男性A.Uのデータでは、以前の確率文法では第2位にランクされていた音韻のうち、2つが改良した確率文法では第1位にランクされたことがわかる。同様に女性S.Uデータでは、以前の確率文法での第2位から、改良した確率文法での第1位にランクアップされたものが3つあることがわかる。また、男性H.Tデータでは、第1位にランクされたデータが改良した確率文法の方が、以前の確率文法より3つ多くなっている。男性N.Mデータでは、第1位にランクされたデータが改良した確率文法の方が、以前より1つ多くなっている。平均して第1位での認識率は、以前の92.1%から92.9%に向上した。

表5.3 男性A.Uデータの音声認識率

rank	HMM-LRのみ	以前の確率文法	改良した確率文法
1	88.2% (246)	91.4% (255)	92.1% (257)
2	98.6% (29)	98.2% (19)	98.2% (17)
3	99.3% (2)	99.6% (4)	99.6% (4)
4	99.3% (0)	99.6% (0)	99.6% (0)
5	99.3% (0)	99.6% (0)	99.6% (0)

※カッコ内は正しく認識した音韻がランクされている数

表5.4 女性S.Uデータの音声認識率

rank	HMM-LRのみ	以前の確率文法	改良した確率文法
1	91.7% (255)	93.9% (261)	95.0% (264)
2	97.1% (15)	99.3% (15)	99.3% (12)
3	99.3% (6)	99.6% (1)	99.6% (1)
4	100.0% (2)	100.0% (1)	100.0% (1)
5	100.0% (0)	100.0% (1)	100.0% (0)

※カッコ内は正しく認識した音韻がランクされている数
※データ数は、都合により278文節

表5.5 男性H.Tデータの音声認識率

rank	HMM-LRのみ	以前の確率文法	確率推定した文法
1	88.5% (246)	92.1% (256)	93.2% (259)
2	95.7% (20)	95.3% (9)	96.4% (9)
3	98.9% (9)	98.6% (9)	98.9% (7)
4	98.9% (0)	98.6% (0)	98.9% (0)
5	98.9% (0)	98.6% (0)	98.9% (0)

※カッコ内は正しく認識した音韻がランクされている数
※データ数は、都合により278文節

表5.6 男性N.Mデータの音声認識率

rank	HMM-LRのみ	以前の確率文法	確率推定した文法
1	89.6% (249)	90.9% (251)	91.3% (252)
2	94.2% (13)	95.7% (13)	96.0% (13)
3	96.8% (7)	97.5% (5)	97.5% (4)
4	97.8% (3)	98.6% (3)	98.6% (3)
5	98.6% (2)	98.6% (0)	98.6% (0)

※カッコ内は正しく認識した音韻がランクされている数
※データ数は、都合により HMM-LRのみ :278文節
確率文法 :276文節

表5.7 全データの平均音声認識率

rank	HMM-LRのみ	以前の確率文法	確率推定した文法
1	89.5%	92.1%	92.9%
2	96.4%	97.1%	97.5%
3	98.6%	98.8%	98.9%
4	99.0%	99.2%	99.3%
5	99.2%	99.2%	99.3%

※データ数は、 HMM-LRのみ :1,113文節
確率文法 :1,111文節

6. 考察

表5.3～5.7より、確率推定方法を改良した文法を用いることによって音声認識率が少しだけ向上することがわかる。なぜ、少しだけ向上するかについて、以下のように考察できる。

以前の推定方法と、改良した推定方法の違いは、解析木の出現に関しての相対確率を考慮するかしないかである。このちがいによる推定された確率の差は、データがあいまいであるほど大きい。しかし、実験の対象である日本語のデーのあいまい度は、1.12(文献[3]による)と小さいため、確率の差は以前の確率文法とそれほど変わらない。これは、複雑度がほとんど変わっていないことからもわかる。したがって、認識率は、少しだけ向上する。

7. むすび

確率推定方法を改良して確率LR構文解析表の確率推定を行った。また、この確率LR構文解析表を用いてHMM-LR音声認識システムで音声認識実験を行い、認識率が以前より少しだけ向上することを確認した。

8. 謝辞

研究の機会を与えていただいたATR自動翻訳電話研究所の樽松明社長に深謝します。また、ATR自動翻訳電話研究所の皆様に感謝します。

9. 参考文献

[1] 北,川端,斎藤:「HMM音韻認識と予測LRパーザを用いた文節認識」,信学技報SP 88 - 88 (1988年10月).

[2] K.Kita : "GLR Parsing in Hidden Markov Model", In "Generalized LR Parsing", ed. Tomita, to appear.

[3] 北,川端,花沢:「HMM - LR音声認識システムにおける統計的言語情報の利用」,信学技報SP 89 - 109 (1990年1月).

[4] T.Fujisaki, F.Jelinek, J.Cocke, E.Black, T.Nishino : "A Probabilistic Parsing Method for Sentence Disambiguation", International Parsing Workshop '89.

[5] 藤崎:「確率的言語処理へのアプローチ」,自然言語処理研究会41 - 6 (1984年1月).

[6] F.Jelinek, R.L.Mercer, L.R.Bahl : "Continuous Speech Recognitions : Statistical Methods", Proc. of the IEEE, Vol. 64, No. 4, 1976.

[7] 中川:「確率モデルによる音声認識」,電子情報通信学会 (1988)

[8] S.Roucos : "Measuring Perplexity of Language Models Used in Speech Recognizers", internal memo.