5H－6

係り受け文法による確率文脈自由文法を用いた言語モデルの検討

柳沼正宣　伊藤彰則　加藤正治　好田正紀
（山形大・工）

1．はじめに

現在,最もよく用いられる言語モデルにN-gramがあるが,自然言語の分析にはよく文脈自由文法に基づいた構文解析が用いられる．このような研究の一つとして確率文脈自由文法（Stochastic Context Free Grammar：SCFG）に関する研究がある．LariとYoung[1],[2]は,確率文脈自由文法の学習を行うInside-Outsideアルゴリズムを提案している．文献[4]ではSCFGの学習を試みていたがN-gramの性能に及ばなかった．また,このアルゴリズムは非終端記号数の３乗に比例して計算時間がかかってしまうという欠点がある．そこで本研究では、Inside-Outsideアルゴリズムに後述する係り受け文法を組み込むことで,性能を落とさずに計算時間を短縮する方法を提案する．

2．確率文脈自由文法（SCFG）

2.1 SCFG

文脈自由文法に確率を付与したものが確率文脈自由文法（SCFG）である．SCFGでは、それぞれの生成規則に確率が与えられる．すなわち,

$$\{\alpha\rightarrow\beta, P(\alpha\rightarrow\beta)\}\in P,\ \alpha\in N,\ \beta\in (N\cup T)^*$$

$$\sum_{\beta} P(\alpha\rightarrow\beta) = 1$$

（N:非終端記号の集合　T:終端記号の集合）

確率文脈自由文法で生成される単語列には,生成確率が定義される．単語列の生成確率は,その生成に用いられた規則の確率の積で表わされる．もし導出にあいまいさがある場合には,複数の導出による確率の和が単語列の確率になる．SCFGの推定には,後述するInside-Outsideアルゴリズムが用いられる。

2.2 Inside-Outsideアルゴリズム

Inside-Outsideアルゴリズムで学習をするためには,生成規則を以下に示すChomsky標準形で表現する必要がある．

$$\alpha\rightarrow\beta\gamma \quad \alpha,\beta,\gamma\in N$$

$$\alpha\rightarrow a \quad \alpha\in N,\ a\in T$$

（N:非終端記号の集合，T:終端記号の集合）

入力テキストを $w_1w_2\cdots w_T$ とする．また,上記の２つのタイプの規則のうち，$\alpha\rightarrow\beta\gamma$ の適用確率を $a(\beta,\gamma|\alpha)$ とし、$\alpha\rightarrow a$ の形の規則が適用される確率を $b(a|\alpha)$ とする．さらに,単語列の i 番目から j 番目までが非終端記号 α に内側から置き換わる確率（Inside確率）を $e(i,j|\alpha)$,外側から置き換わる確率（Outside確率）を $f(i,j|\alpha)$ とする．このとき,

$$g(i,k,j;\alpha\rightarrow\beta\gamma) = e(i,k|\beta)e(k+1,j|\gamma)a(\beta,\gamma|\alpha)f(i,j|\alpha) \quad (1)$$

$$a'(\beta,\gamma|\alpha) = \frac{\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\sum_{k=i}^{j-1} g(i,k,j;\alpha\rightarrow\beta\gamma)}{e(1,T|S)} \quad (2)$$

$$b'(a|\alpha) = \frac{\sum_{t;w_t=a} b(a|\alpha)f(t,t|\alpha)}{e(1,T|S)} \quad (3)$$

によりa,bを再推定することができる．

3．係り受け文法を用いたSCFG

係り受け文法とは,文節間の修飾関係によって文構造を規定しようとするもので,文節x が文節y を修飾するとき、x はy に係り,yはx を受けるという[3]．係り受けは日本語特有の規則である．

本研究において,係り受け文法を用いるためInside-Outsideアルゴリズムの改良を試みた．まず,単語を実質語と機能語に分ける．機能語とは，”助詞”,”助動詞”,”接尾語”,”語尾”のことをいう．機能語以外を実質語とする．そしてChomsky 標準形の代りに以下の規則を適用する．

(1) $\alpha\rightarrow\beta\alpha$　　$\alpha,\beta\in N$

(2) $\alpha\rightarrow b$　　$\alpha\in N$　$b\in T_c$

(3) $\alpha\rightarrow\beta a$　$\alpha,\beta\in N$　$a\in T_f$

（但し，T_f は機能語の集合，T_c は実質語の集合）

規則(1)は文節間文法を表現し,規則(2)は１つの非終端記号が１つの終端記号（実質語）を生成する規則である．規則(3)は文節内文法を表現している．

(1),(2),(3)の規則が生成される確率をそれぞれ $a(\beta|\alpha)$, $b(b|\alpha)$, $c(\beta,a|\alpha)$ とすると,再推定式は以下のように表わせる．

$$g(i,k,j;\alpha\rightarrow\beta\alpha) = e(i,k|\beta)e(k+1,j|\alpha)a(\beta|\alpha)f(i,j|\alpha) \quad (4)$$

$$a'(\beta|\alpha) = \frac{\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\sum_{k=i}^{j-1} g(i,k,j;\alpha\rightarrow\beta\alpha)}{e(1,T|S)} \quad (5)$$

A Study of Language Modeling using Stochastic Context Free Grammar with Dependency Grammar
Masanobu Yaginuma,Akinori Ito,Masaharu Katoh,Masaki Kohda;Yamagata University

$$b'(b\,|\,\alpha) = \frac{\sum_{t;w_t=b} b(b|\alpha) f(t,t\,|\,\alpha)}{e(1,T|S)} \quad (6)$$

$$c'(\beta,a|\,\alpha) = \frac{\sum_{i;w_i=a}^{T} \sum_{r=1}^{i-1} e(r,i-1|\beta) f(r,i|\alpha) c(\beta,a|\alpha)}{e(1,T|S)} \quad (7)$$

このアルゴリズムを用いると,$O(|N|^2)$ (非終端記号数の２乗)での計算時間で学習することができる.

4．実験

4.1 実験条件

EDRコーパス先頭2100文のうち前半100文が評価用、後半2000文を学習用に使用する．学習テキスト2000文の単語数は49910,実質語の種類数は8042,機能語の種類数は415である．評価テキスト100文の単語数は2583,実質語の種類数は704,機能語の種類数は104である.

学習テキストにおいて、頻度１以下の単語を未知語として特定の単語(UNK)に置き換える．このときの学習テキストの未知語の数は5480種類（実質語5321,機能語159)である．評価の方法は,タスクperplexityを用いる． 使用計算機はPeutium90MHzのPC互換機(Linux)である.

4.2 実験結果

まず,図１に係り受けを用いたSCFGとOriginalのSCFG の計算時間の比較を行なっている縦軸は１回の再推定に必要なCPU時間（秒）,横軸は非終端記号数である.

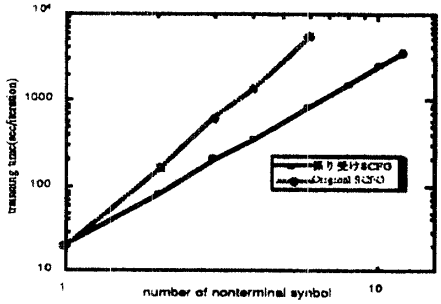

図１ 非終端記号数と学習時間の関係

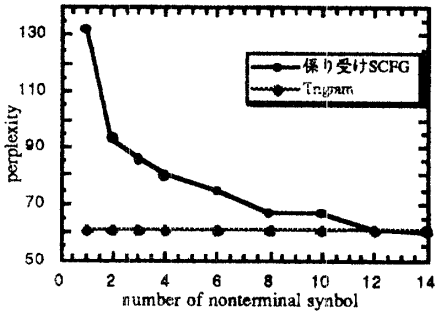

図２ 非終端記号数とPerprexityの関係

表１ 各モデルのパラメータ数

Models	number of parameters
係り受けSCFG(12)	53544
Original SCFG(12)	37476
Bi-gram	50586
Tri-gram	74199

注. () 内の数字は非終端記号数

前者は非終端記号数の３乗に比例していて,後者は２乗に比例していることが分かる．つまり係り受けを用いることで,実現可能な時間で計算できるようになる.

図２は係り受けSCFGの性能を示している．非終端記号数12のときのperplexityは約60.3となっている．Trigram(CMUツールキット使用[5])のperplexityは60.4であった.

表１は各モデルのパラメータ数の比較である．係り受けSCFGのパラメータ数はTrigramのパラメータ数よりもかなり少ない．また、Original SCFG はパラメータ数が少ないが,計算時間がかかるため非終端記号数12では学習が困難である.

6．まとめ

係り受けを用いることにより計算時間の削減が可能になり,Trigramと同等以上の性能を示した．今後の課題として,パラメータの初期値の取り方について検討し,計算時間の削減と性能の向上を目指す.

参考文献

[1] K.Lari,S.J.Young:The estimation of stochastic context-free grammars using the Inside-Outside algorithm:Computer Speech and Language,4, 35-56,PP1990

[2] K.Lari,S.J.Young: Application of stochastic context-free grammars using the Inside-Outside algorithm: Computer Speech and Language, 5,PP.237-257,PP.1991

[3] 中川聖一: 確率モデルによる音声認識: (社) 電子情報通信学会(1988)

[4] Min Zhou:A Study on Stochastic Models for Spoken Language : Ph.D thesis,Toyohashi University of Technology, 1996

[5] Ronald Rosenfeld: The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evalution:Proceedings of the Spoken Language Systems Technology Workshop,PP.47~50,1995