確率文脈自由文法におけるk best parserにおける妥当なkの検討

渥美 清隆　　増山 繁
atsumi@smlab.tutkie.tut.ac.jp,　masuyama@tutkie.tut.ac.jp
豊橋技術科学大学　知識情報工学系

1 はじめに

自然言語の分析のために, 文脈自由文法を利用した構文解析を利用することが多い. しかし, 文脈自由文法には曖昧性が含まれることが多く, この場合1つの入力列に対して, 複数の解析結果を構文解析が出力することになる. 複数の解析結果があまり多くならなければ, 意味解析などに選択を委ねることも可能であるが, 非常に多くの解析結果を出力することになった場合には, 意味解析だけで選択することは難しく, それ以外の簡便な選択方法が必要になる.

そこで, 解析結果に対する確率値を付与し, その確率値の大きさによって解析結果を絞ることが考えられる. これは確率文脈自由文法[1]によって実現できるが, 解析結果をどの程度絞ることができるのかを議論した論文は見つけることができなかった.

本論文では, 解析木に付与される確率についていくつかの仮定を置き, 解析木の確率値の出現確率の密度関数と重み付きの密度関数を計算することにより, 解析結果をどの程度絞り込むことができるのかを検討する.

2 確率文脈自由文法[1]の定義

ここでは, 確率文脈自由文法$SCFG$ (Stochastic Context-Free Grammar)と解析木への確率の付与の仕方について定義する.

2.1 確率文脈自由文法の定義

確率文脈自由文法は$SCFG = (N, \Sigma, S, P, p)$の5つ組で表現され, それぞれ以下のように定義する.

- Nは非終端記号の集合
- Σは終端記号の集合
- Sは出発記号
- Pは導出規則の集合
- pは$P \to [0,1]$の写像

導出規則は

$$A \underset{p(A \to \beta)}{\to} \beta$$

$$A \in N,\ \beta \in (N \cup \Sigma)^*$$

で表現される. 各導出規則に確率値pが割り当てられていることを除けば, 通常の一般文脈自由文法と全く同じである. $p(A \to \beta)$は導出規則$A \to \beta$を使うときの非終端記号Aに対する条件付き確率である. Aから導出される規則がk個あり, それぞれの導出規則を$A \to \beta_i$, $1 \le i \le k$とするとき, $\sum_{i=1}^{k} p(A \to \beta_i)$は1で正規化されている.

2.2 確率文脈自由文法から得られる解析木の確率の定義

確率文脈自由文法を用いた構文解析は一般の文脈自由文法の構文解析と何等変わることはない. しかし, 構文解析を行った後, それぞれの解析木に確率値を付与するために再計算を行う. 各解析木Tはいくつかの導出規則$P_1, \cdots, P_i$から成り立っており, それぞれの導出規則には確率値$p(P_1), \cdots, p(P_i)$が付与されている. これらの確率値から, 以下のような計算をする.

$$p(T) = \prod_{j=1}^{i} p(P_j)$$

この計算により各解析木の確率値が与えられる.

3 解析木に付与される確率の近似

我々が興味を持っている情報は, 構文解析から出力される多数の解析木の内, 確率的にみて意味のある解析木がどのくらい存在しているかである.

導出規則は解析木の節点に付与される非終端記号に従う. この導出規則の使用頻度や使用位置は入力列に応じた複雑な振る舞いをするため, 具体的な入力列が定まらない限りどのような導出規則が使用されるかは特定できない. このため, 構文解析が出力する解析木に付与される確率値の情報, 例えばaから1までの

範囲の確率値を持つ解析木が, 各入力列の平均でいくつ存在するのか等の情報を得ることを困難にしている.

解析木に付与される確率値の振る舞いを複雑にしている理由を整理すると,

1. 一般文脈自由文法では解析木に使用される導出規則の数が特定できない.
2. 導出規則の使用頻度や位置は入力列に依存する.
3. 導出規則に付与されている確率値の分布情報が離散的である.

である. これらの問題を解決するために以下の3つの仮定を置く.

1. 文法はChomsky標準形で書かれている.
2. それぞれの導出規則が使用される頻度は一様であり, 位置は互いに独立である.
3. 導出規則に付与されている確率値の出現確率を示す確率変数は微分可能な密度関数で表現できる.

1つ目の仮定の目的は1つの入力列から得られる解析木の使用する導出規則の数を固定するためである. Chomsky標準形で書かれた文法を利用して構文解析を行うと, 入力列の長さがnのとき, 解析木に使われる導出規則の数mは$m = 2n - 1$個に固定することができる.

2つ目の仮定の目的は1次従属による複雑な振る舞いをする解析木に付与される確率値の計算を, 単純な振る舞いに変換するためである. ここで単純化計算法の概略を説明する. 導出規則に付与されている確率値をボールに1つづつ書き写し, それを大きな袋に入れることを想定する. もし, 入力列の長さがnであるならば$m = 2n - 1$個のボールを復元抽出で袋から取り出し, ボールに記録されている数字を書き取る. この数字の全ての積がその入力列から得られるある解析木に付与される確率値であるとする. この試行を何回か繰り返したときに得られる密度関数を解析木に付与される確率値の出現確率の密度関数とする.

3つ目の仮定の目的は, 離散的な振る舞いから連続的な振る舞いに換えることにより, 微積分などの手段を利用可能にすることである.

3.1 評価方法の定義

導出規則に割り当てられた確率値の出現確率が確率変数X_1に従うとする. この確率変数の密度が$f_1(x)$で表現できるとき, m個の導出規則を使った解析木に付与される確率値の出現確率は確率変数X_mに従う. この確率変数の密度は

$$f_m(x) = \int_x^1 f_1(\frac{x}{y}) f_{m-1}(y) \frac{1}{|y|} dy \qquad (1)$$

と定義する. これはm個の独立した確率変数から得られる確率の積と同一である[2]. また, $xf_m(x)$は重み付きの密度関数であるとするとき,

$$F_{m,a}(x) = \int_a^1 x f_m(x) dx \qquad (2)$$

はaから1までの確率値が付与された解析木の重みの合計を表わす.

(2)式においてaから1までの解析木の重みの合計がrに達するようなaを求めれば, 解析木に付与される確率の大きい順に解析木を選択するとき, いくつまで選択すればよいのかの目安にすることができる. このaを求める式として以下を定義する.

$$\max_a \left\{ \frac{F_{m,a}(x)}{F_{m,0}(x)} \geq r \right\} \qquad (3)$$

そして, (3)式で求めたaから, 全体の解析木の数と比較してどのくらいの割合で解析木を取り出せばよいかは, 次の式に従う.

$$RNT_{m,a}(x) = \int_a^1 f_m(x) dx \qquad (4)$$

この(4)式から得られる数がmと共にどのように変化するかを考察する. 以下の節では導出規則に付与される確率値の出現確率の密度関数について具体的な連続関数を想定して考察する.

3.2 密度関数が一様関数の場合

導出規則に付与される確率値の出現確率の密度関数が$f_1(x) = 1, 0 \leq x \leq 1$に従うとき, つまり確率値が0.1, 0.2, $\cdots$, 1.0のように等間隔に存在する場合を考える. このとき$f_2(x)$は(1)式から

$$f_2(x) = \int_x^1 f_1(\frac{x}{y}) f_1(y) \frac{1}{|y|} dy = -\log x$$

が得られる. この式が持つaから1までの重み付き密度関数の積分値は

$$F_{2,a}(x) = \int_a^1 x f_2(x) dx = \frac{1}{4} + \frac{a^2}{2}(\log a - \frac{1}{2})$$

となる．$a=0$のとき，この式の値は1/4になるので，$r=0.8$としてこれを(3)式に代入すれば，

$$\max_a \left\{ \frac{\frac{1}{4}+\frac{a^2}{2}(\log a-\frac{1}{2})}{\frac{1}{4}} \geq 0.8 \right\}$$

となる．この式からaは約0.22と求めることができる．そしてこのaの値を(4)式に代入すると，

$$RNT_{2,0.22}(x) = -\int_{0.22}^{1} \log x dx = 0.447$$

を得る．つまり，構文解析から出力される解析木の内，解析木に付与された確率が大きい順に，解析木の総数の44.7%を選択すればよいことになる．

さて，この議論が実際の導出規則が持つ確率値の振る舞いと一致するかどうかの簡単な検証を行う．　まず，導出規則に付与された確率値は$\hat{f}_1 = \{0.1,\ 0.2, \cdots, 1.0\}$，$|\hat{f}_1| = 10$であったとする．　導出規則を独立に2つ選択し，それぞれが持つ確率値の積を求めたとき，$\hat{f}_2 = \{0.01,\ 0.02,\ 0.02,\ 0.03, \cdots, 0.81,\ 0.9,\ 0.9,\ 1.0\}$，$|\hat{f}_2| = 100$を得る．このf̂2の集合から大きい順に49個を選択すると全体の重みの80.3%に達する．(4)式に対応する計算では$49/100 = 0.49$となるので，連続関数上での議論とほぼ一致するものと考える．連続関数上の議論と離散関数上での議論の差は最初の導出規則に付与される確率値の密度関数の微妙な差から生まれていることが考えられる．

次に入力列の長さを制限しない一般の場合について検討する．$f_2(x)$，$f_3(x), \cdots$と順に求めていくと，

$$f_m(x) = \frac{(-1)^{m-1}}{(m-1)!} \log^{m-1} x$$

を得ることができる．この式を(2)式に代入すると，

$$\begin{aligned}
F_{m,a}(x) &= \int_a^1 x f_m(x) dx \\
&= \frac{(-1)^{m-1}}{(m-1)!} \left\{ \left[\frac{x^2 \log^{m-1} x}{2} \right]_a^1 \right. \\
&\quad \left. -\frac{m-1}{2} \int_a^1 x \log m - 2x dx \right\} \\
&= \frac{(-1)^{m-1}}{(m-1)!} \left[\frac{x^2 \log^{m-1} x}{2} \right. \\
&\quad -\frac{m-1}{2} \left(\frac{x^2 \log^{m-2} x}{2} \right. \\
&\quad -\frac{m-2}{2} \left(\frac{x^2 \log^{m-3} x}{2} \right. \\
&\quad \vdots \\
&\quad -\frac{2}{2} \left(\frac{x^2 \log x}{2} \right. \\
&\quad \left. \left. \left. -\frac{1}{2} \left(\frac{x^2}{2} \right) \cdots \right) \right] \right]_a^1 \qquad (5)
\end{aligned}$$

となる．また，(4)式に代入すると，

$$\begin{aligned}
RNT_{m,a}(x) &= \int_a^1 f_m(x) dx \\
&= \left[\frac{(-1)^{m-1}}{(m-1)!} x \log^{m-1} x \right. \\
&\quad + \frac{(-1)^{m-2}}{(m-2)!} x \log^{m-2} x \\
&\quad \vdots \\
&\quad + \frac{(-1)^1}{1!} x \log x \\
&\quad \left. + \frac{(-1)^0}{0!} x \right]_a^1 \qquad (6)
\end{aligned}$$

となる．本来なら(5)式を(2)式に代入し，一様関数におけるaを求めるnの関数式を作り，その関数式を(4)式に代入するべきであるが，単純な式にまとめることが難しいため，コンピュータ上で$m=1 \sim 100$までについてaを計算し，その結果の一部を表1にまとめた．また，ここで計算されたそれぞれのaの値を(6)式に代入した結果の一部を表2にまとめた．表中のrは確率の大きい順に解析木を選択した場合の重みと解析木の重み全体に対する比率である．

表1，2から入力列長nの増加量に対して，どちらも指数的に減少していることが分かる．特に，解析木の選択割合の指数的な減少は，入力列長に応じて指数的に増加する解析木を抑制する効果が期待できる．

3.3　分布関数が一様分布以外の場合

導出規則に付与される確率値の出現確率の密度関数の一般的な性質として，密度関数の平均が0.5を超えることがほとんど有り得ない，ということである．なぜならば，確率値1を持つ導出規則を除いた導出規則に付与される確率値の平均は必ず0.5以下であり，確率値1を持つ導出規則は通常ごく僅かであると考えられるからである．また，導出規則の確率値の出現確率の密度関数の平均がμであるとき，解析木の確率値の出現確率の密度関数の平均は入力列長がnであるならばμ^{2n-1}になる．そして，解析木の確率値の出現確率の密度関数を導出規則の組み合わせとして考えるとき，

組み合わせたときの調整パラメータとして働く$1/|y|$の項がlogの多項式に変化するため, nがある程度大きくなるとlogの多項式が他の項に比べて支配的になる. このため, 導出規則に付与される確率値の出現確率の密度関数がどのような関数であろうと, 一様分布の場合とあまり変わらず, nが大きくなるに従って, 解析木の選択割合は指数的に減少することが期待できる.

4 まとめ

本論文では, 確率文脈自由文法に基づく構文解析をした結果の解析木に付与される確率値の振る舞いについて, 3節の3つの仮定の上で, 以下のことを明らかにした.

1. 確率値が非常に小さな解析木が大多数を占め, 確率的に意味のある解析木は非常に僅かである.
2. 入力列が長くなると, 全体の解析木の数に比べて確率的に意味のある解析木は指数的に少なくなる.
3. 一様関数だけでなく, 導出規則に付与される確率値の密度関数がいかなる関数であっても, 上記2つの性質を期待することができる.

この結果, 入力列長に応じて解析木の数が非常に増加するような文法規則でも確率文脈自由文法上では解析木の数を抑制することが期待できる. 特に入力列に対する解析木の増加割合が解析木の選択割合の指数的減少と同期しているか, またはそれよりも遅い増加割合であるときは解析木は確率の高い順に高々$O(n)$個程度選択すれば十分である.

この結果を導くために3節で3つの仮定を置いている. 今後の課題として, この仮定の妥当性を検証するために実際の確率文脈自由文法に基づく構文解析による実験が必要である.

表 1: 一様関数の場合の各入力列長n(導出規則の数m)に対するaの値

$n(m)$	$r=0.2$	$r=0.5$	$r=0.8$
2(3)	0.464	0.262	0.118
5(9)	0.0402	0.0131	0.00338
10(19)	0.000484	8.84e-5	1.28e-5
15(29)	5.03e-6	5.96e-7	5.57e-8
20(39)	4.88e-8	4.01e-9	2.61e-10
25(49)	4.52e-10	2.70e-11	1.28e-12
30(59)	4.06e-12	1.82e-13	6.45e-15
35(69)	3.57e-14	1.23e-15	3.34e-17
40(79)	3.07e-16	8.27e-18	1.76e-19
45(89)	2.60e-18	5.57e-20	9.42e-22
50(99)	2.18e-20	3.76e-22	5.11e-24

表 2: 一様関数の場合の各入力列長n(導出規則の数m)に対する解析木の選択割合

$n(m)$	$r=0.2$	$r=0.5$	$r=0.8$
2(3)	0.0429	0.151	0.361
5(9)	0.00587	0.0331	0.122
10(19)	0.000375	0.00357	0.0219
15(29)	3.02e-5	0.000431	0.00395
20(39)	2.71e-6	5.48e-5	0.000705
25(49)	2.61e-7	7.16e-6	0.000124
30(59)	2.63e-8	9.52e-7	2.18e-5
35(69)	2.75e-9	1.28e-7	3.78e-6
40(79)	2.94e-10	1.74e-8	6.52e-7
45(89)	3.22e-11	2.39e-9	1.12e-7
50(99)	3.59e-12	3.29e-10	1.90e-8

参考文献

[1] 中川: "確率モデルによる音声認識", 信学会(1988).

[2] A.M.Mood, F.A.Graybill and D.C.Boes: "Introduction to the Theory of Statistics", *McGraw-Hill* (1974). 和訳大石: 統計学入門(上) マグロウヒル好学社(1978).