統計情報に基づくチャートパーザの制御

2Q－5

加藤　恒昭
NTT情報通信網研究所

James F. Allen
Rochester大学計算機科学科

はじめに

解析の対象となる分野に現われる文とその統語構造を大量に収集したコーパスから解析の尤もらしさに関する統計情報を収集し、その情報を用いてパーザを制御することにより、正しい解を高速に得ようとする試みが数多くなされている[1]-[3]。本稿では、その様な試みのひとつとして、素性に基づく文法に対するチャートパーザを、コーパスから得られた統計情報を利用して制御する方式について提案する。

本方式は解析途中で得られる弧に得点を与え、得点の高い弧を優先的に処理するという一般的なアジェンダ制御であるが、得点の与え方に特徴がある。つまり、部分統語構造や規則にではなく弧自体にコーパス中での出現頻度に対応した得点を与える。しかもその得点は自分を構成する部分構造の尤もらしさから独立である。極めて小規模ではあるが、初期的な実験の結果、本方式の有効性が確認されている。

文法とパーザ

利用した文法の枠組みはRochester大学において行なわれているTRAINSプロジェクト[4]の一環として研究されているもので、GPSG流の素性に基づく記述が用いられている。統語カテゴリに相当するものは素性の集まりとして定義され、各素性は素性木の節点として意味づけられる。この素性木において娘達は母節点の相互排他的で網羅的な部分であり、ある節点はその子孫および祖先と単一化可能である。例として動詞の型を表現する素性木を図1に示す。

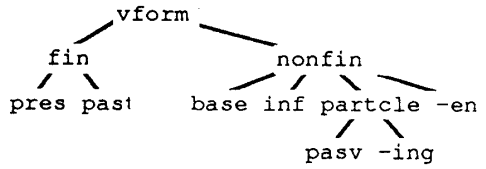

図1　素性木の例（動詞の型）

各文法規則は例えば図2に示すような形をしている。統語部は前述の素性により記述される。素性の伝播は規則中の記述に加え、主辞素性原理等によっても定められている。統語部と一対一に対応する意味部は

```
(S_N2+V2 ;; [we] [make OJ]
 :SYNRULE ((S full-decl) (N 2bar nom) (V 2bar))
 :SEMRULE (i 1 2)
 :HEADS   (2)
 :CONSTRAINTS (((1 AGR) 2 AGR)) )

(V2_V+N2+P2path ;; [get] [a boxcar] [to Bath]
 :SYNRULE ((V 2bar)
      (V _N2_P2path) (N 2bar acc) (P 2bar path))
 :SEMRULE (f (f adv-a 3) (p 1 2))
 :HEADS   (1) )
```

図2　文法規則の例

Episodic Logicからなる論理形式を生成する。

このような枠組みをチャートパーザで扱うために、ここでは、これを拡張CFGと見做すという方針をとった。つまり、幾つかの中心的な素性木、X素性、Barレベル、SUBCAT等の値を合成して得られる相互排他的なシンボルを統語カテゴリとして扱うことで骨格となるCFGを構成し、それ以外の素性は拡張項に相当するものとした。例えば、図2の文法規則は各々次の様に表現できる。ここで、S,N2,V2,V_N2_P2path等が中心的な素性から合成されたCFGカテゴリである。

```
S(full-decl) → N2(nom) V2
V2 → V_N2_P2path N2(acc) P2(path)
```

この扱いに基づいてEarleyタイプの予測を行なう下降型チャートパーザ[5]を以下の様に実現した。

- 予測弧生成における弧の重複管理と到達可能性に基づく枝刈りは骨格となるCFGについて行なう。
- 素性の単一化は活性弧と非活性弧の合成時(Fundamantal rule[5]適用時)にのみ行ない、素性の情報は上方向にのみ伝播させる。

このような実現は、パーザの予測力と重複管理の複雑さの適切な妥協点である。また、統語カテゴリ構成のために用いられた素性については素性木の利点が活かせないことが問題となるが、これについては簡単な前処理により回避できる。

統計情報による制御

前述した下降型チャートパーザに対してアジェンダ制御を行なったわけであるが、弧の得点として式(1)を利用した。つまり、図3に示す始点m、終点n、ドット付き規則$\alpha \rightarrow \psi \bullet \varphi$なる弧の、入力S解析時における得点は、弧の終点近辺の語トライグラムを条件とした、ドット付き規則(不完全な部分構造)の出現確率から求められる。この値は式(2)のように統語構造付きコー

Stochastic Control of a Chart Parser
Tsuneaki Kato
NTT Network Information Systems Labs.
1-2356 Take, Yokosuka-shi, Kanagawa, 238-03

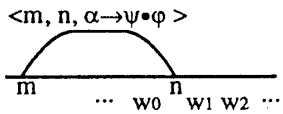

図3　弧と語トライグラム

パスを用いて、正しい構造を導いた弧やトライグラムの出現数を計算することで近似できる。

$$SCORE(<m, n, \alpha\rightarrow\psi\bullet\varphi>, S) = P(\alpha\rightarrow\psi\bullet\varphi \mid w_0w_1w_2) \quad (1)$$

ここでTrgrm(S, n) = $w_0w_1w_2$、つまり、Sのn番目の単語を中心とする語トライグラムは$w_0w_1w_2$。

$$P(\alpha\rightarrow\psi\bullet\varphi \mid w_0w_1w_2) \cong \frac{N(<?, i, \alpha\rightarrow\psi\bullet\varphi>, S)}{N(\ Trgrm(S, i) = w_0w_1w_2\ )} \quad (2)$$

なお、語彙規則から導かれる弧については、トライグラムを用いた品詞推定と類似した手法により得点付けを行なっている。また、後述の実験においては、直接には語トライグラムを利用せず、カテゴリトライグラムを用いて近似し、更に学習量の不足を補うためバイグラムを利用した補完を行なっている。

従来の方法が、部分構造(非活性弧)や文法規則に得点を与え、それらを基に、例えば積や相乗平均を利用して不完全な部分構造(活性弧)の得点を導いていた[2]、もしくは活性弧には得点を与えなかった[3]のに対し、本手法では、活性弧に直接、コーパスでの出現頻度に基づく得点を与えている。その結果、ある構造の得点はその下位部分構造の得点から独立となり、PCFG[1]等を利用したときに問題となる、より大きな構造ほど低い得点が与えられ幅優先探索に近い制御に陥ってしまうという事態は完全に回避される。更に、ある構造の尤もらしさの文脈依存性は、その構造の右端近辺の語トライグラムを出現頻度の文脈としていることで表現されている。

初期実験

TRAINSの分野で収録された対話文を用いて小規模な初期実験を行った。260文を対象とし、言い誤りの修正や適当な句読点を付与を人手で行った後、それら全てを受理する文法を作成し、正しい解析結果を人間が選択して文と統語構造のペアを収集した。得られた文法は、文法規則数2090(うち句構造規則246)、素性数326、素性木数38、骨格となるCFGのカテゴリ数74である。また、コーパス中の文は、平均単語数11.6、解析に必要な弧の数の平均は45.7で、95文(37%)は曖昧性がなく、74文(28%)は10以上の統語構造候補を持つ。この文法とコーパスを用いて2種類の実験を行なった。なお、制御において利用したカテゴリは骨格であるCFGのカテゴリを基本としたが、wh素性の有無を区別し、助詞については単語そのものを用いている。

実験1：実験対象文260文を5つのセットに分割し、4つのセットで学習を行い(つまり統計情報をそれらから計算し)、残ったひとつのセットを対象に実験を行う。この実験を実験セットを変えて5回行った。

実験2：260文すべてで学習を行ない、同じ260文で実験を行なった。

実験結果を表1に示す。調査した統計量は次の通りで、比較は2種類のPCFG(下位構造からの計算にそれぞれ積と相乗平均を用いている)と行った。

正解数：正しい構造が最初に得られた文の数

弧数：正しい構造が最初に得られた文について、処理した弧の数の平均、括弧内は最少数との比

誤数：誤った構造が最初に得られた文の数

未終了：5000本の弧を処理してひとつの解も得られなかった文の数

表1　実験結果

実験1	正解数	弧数	誤数	未終了
提案手法	190(73%)	353(7.7)	59(23%)	11(4%)
PCFG(積)	154(59%)	1677(36.7)	38(15%)	68(26%)
PCFG(相乗)	157(60%)	855(18.7)	78(30%)	25(10%)
制御無し	134(51%)	1259(27.5)	62(24%)	64(25%)

実験2	正解数	弧数	誤数	未終了
提案手法	255(98%)	83(1.3)	5(2%)	0
PCFG(積)	151(58%)	1636(35.8)	42(16%)	67(26%)
PCFG(相乗)	165(63%)	644(14.1)	78(30%)	17(7%)

考察

初期実験の結果、少ない学習量にもかかわらず、提案手法では、正確度、速度共に比較的よい結果が得られることが確認された。今後の課題としては、文法、解析対象文共に拡充して、実際的な規模での実験を行い、方式の優位性を検証することが挙げられる。また、今回利用した統計量が、数学的にどのような意味を持つのかについても、より厳密な考察が必要である。

参考文献

[1] Chitrao, M.V. and Grishman, R. Statistical parsing of messages. DARPA Speech and NL 1990

[2] Magerman, D. and Weir, C. Efficiency, robustness and accuracy in Picky chart parsing. ACL 1992

[3] Tashiro, T. et al. Efficient chart parsing of speech recognition candidates. ICASSP 1994 (to appear)

[4] Allen, J.F. and Schubert, L.K. The TRAINS project. Univ. of Rochester Tains TR-91-1 1991

[5] Gazdar, G. and Mellish, C. Natural language processing in LISP. Addison-Wesley 1989