文献紹介ゼミ

構文解析にもとづく規則生成と規則集合探索による文脈文法の漸次学習

長岡技術科学大学　自然言語処理研究室
後藤　大明

出典

- 中村 克彦, 保科 明美,構文解析にもとづく規則生成と規則集合探索による文脈自由文法の漸次学習,人工知能学会論文誌, vol 21, pp.371-379, (2006)

- キーワード
- grammatical inference, CLF, incremental learning, bottom-up parsing

まえがき

- 文法推論
  - 正規言語などの比較的制約の強い言語が中心
  - 文脈自由文法を学習する研究は少数

- 文脈自由文法の推論に必要な計算量は非常に多い[Angluin 95, Pitt 93]
- 発見法と探索方式の改良し規則生成の高速化
- 学習モデルを導入し規則集合を生成

まえがき

- 帰納的CYKアルゴリズム
  - 上向き構文解析法　文法学習システムSynapseで有効性を確認[Nakamura 00]

- より複雑な文法を効率よく学習できるよう規則生成方式 ブリッジ法 探索方式 直列探索を導入

文脈自由文法とChomsky標準形

- 文脈自由文法　G=（N,T,P,S）
  - Gは文法、N,T,Pは非終端記号、終端記号、生成規則の有限集合
  - S∈Nは開始記号

- 生成規則
  - $A \rightarrow u,\ A \in N,\ u \in (N \cup T)^{+}$

変形Chomsky標準形

- Chomsky標準形(CNF)
  - 規則$A \rightarrow BC$　及び　$A \rightarrow a$のみで表される

- 変形Chomsky標準形(変形CNF)
  - 規則$A \rightarrow \beta\gamma(\beta,\gamma \in N \cup T)$
  - 長さが1の文字列を含まない
- 終端記号が少ない言語に対して規則集合を少なくできる

拡張Chomsky標準形

- 拡張Chomsky標準形(拡張CNF)
  - 変形CNFにA→Bの形式を追加したもの

規則生成アルゴリズム

- 正例の文字列と規則集合に対して構文解析を行い、失敗した時成功するまで規則を追加する

帰納的CYKアルゴリズム

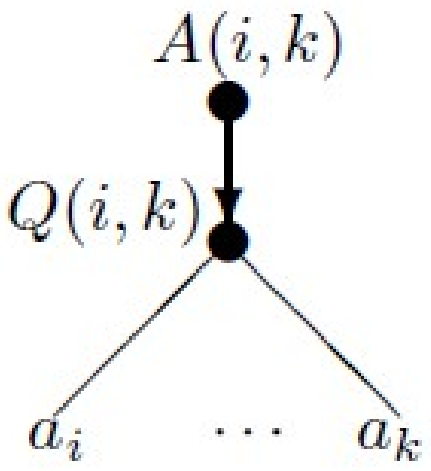
- 上向き構文解析によって開始記号Sからの導出ができないとき規則を追加する

- 規則の生成
  - 1.対βγ∈TS(規則の適用を試みた記号対の集合)を非決定的に選択
  - 2.非終端記号Aを非決定的に選択
  - 3.規則A→ βγを追加
  - 4.再度構文解析

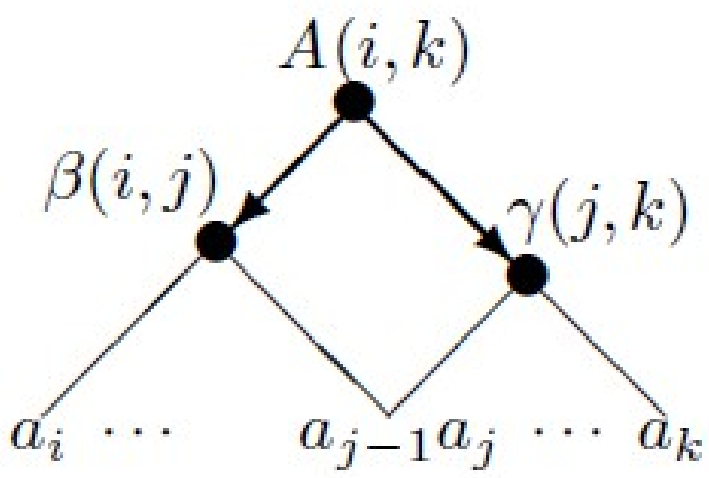
ブリッジ法による規則生成

- 頂点ラベルSを持つ導出木が生成されないとき(1)～(6)の演算を非決定的に行う

(1) $A \to Q$ を生成.

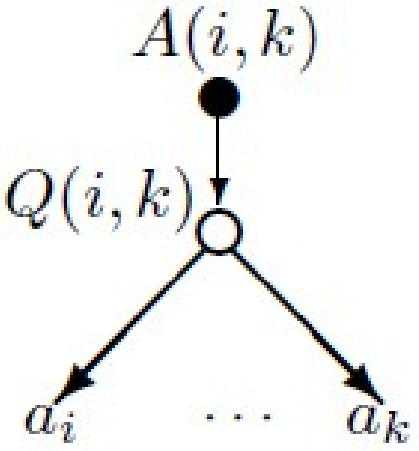


(2) $A \to \beta\gamma$ を生成.

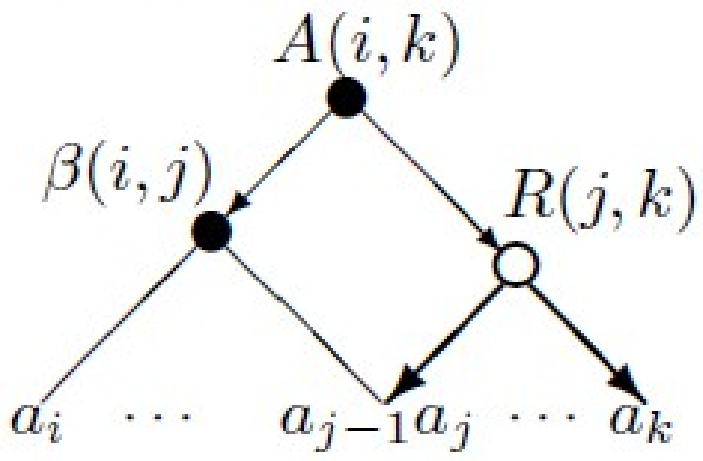


ブリッジ法による規則生成

(3) $A \to Q \in P$ ならば, $Q(i,k)$ 以下の規則を生成.

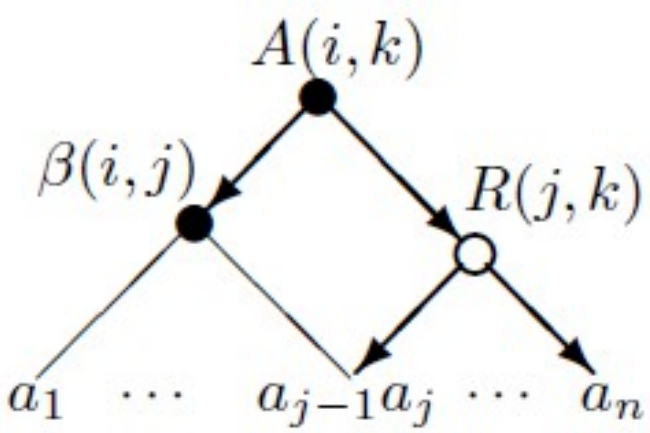


(4) $A \to \beta R \in P$ ならば $R(j,k)$ 以下の規則を生成.

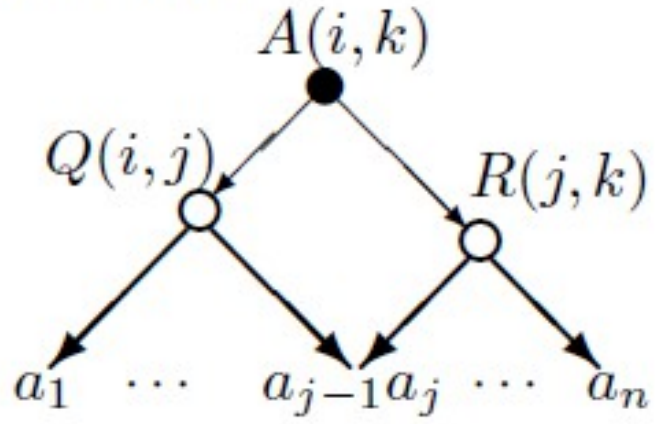


ブリッジ法による規則生成

(5) $A \to \beta R$ を生成，$R(j,k)$ 以下の規則を生成.



(6) $A \to QR \in P$ ならば，$Q(i,j)$ と $R(j,k)$ 以下の規則を生成.



文法推論のための探索方式

- 反復深化によって最小の規則集合を求める
- 最小規則集合探索
  - 規則の上限を設定し、構文解析が成功するまで上限を1づつ増やす

文法推論のための探索方式

- 直列探索方式
- 各正例を導出する最小の規則集合を部分解とし、全ての正例に対する最小規則集合を探索しない

- 2つの探索法で学習された文法が与えられた正例と負例を満足することは数学的帰納法で証明可能

実験方法

- “オートマン, 言語理論, 計算論I”[Hopcrft 79]のCFG作成の練習問題となっている7つの文法を構成する課題のうち5つ

- Synapseに組み込まれた帰納的
  - CYKアルゴリズム・最小規則集合探索(CM)
  - ブリッジ形規則生成・最小規則集合探索(BM)、
  - ブリッジ形規則生成・直列探索(BS)

結果

言語	学習時間(秒) ： 全生成規則数		
	CM	BM	BS
a	0.02:179	0.05:32	0.01:7
b	0.1:10000	0.72:596	0.05:39
c	8786:8×10^{7}	112:2.2×10^{4}	4.0:874
d	85:4.4×10^{4}	1185:2.1×10^{6}	6:7213
e	570:5.7×10^{6}	579:2.4×10^{6}	-

むすび

- ブリッジ法により効率よく規則を生成した
- 直列探索による文法学習は最小規則探索と比べてかなり少ない計算時間の学習が可能

- ただし言語によっては直列探索の計算時間が非常に大きくなる場合がある