語句の分布情報に基づいた多重文脈自由文法の学習

吉仲　亮

科学技術振興機構　ERATO湊離散構造処理系プロジェクト

文　法　推　論

・形式言語のアルゴリズム的学習
・母語獲得メカニズムの数理モデル化
・自然言語処理，生物情報 etc...
・豊かな言語族を合理的な枠組みで効率的に学習
・例からの学習，質問による学習，確率的学習 ...

語句の分布情報に基づく学習

・文を，部分文字列と文脈（接頭辞と接尾辞の対）の合成として捉え，その関係に注目する
・$(u, v) \times w = uwv \in L$?

可代入文脈自由言語

・Clark & Eyraud (2007)

$u_1w_1v_1,\ u_1w_2v_1,\ u_2w_1v_2 \in L \Rightarrow u_2w_2v_2 \in L$

言語レベル	注目する語句分布
正規言語	$u \times v = uv$
文脈自由言語	$(u,v) \times w = uwv$
多重文脈自由言語	$(u_0, u_1, \ldots u_m) \times (v_1, \ldots v_m) = u_0v_1u_1 \ldots v_mu_m$

多重文脈自由文法

・自然言語の表現に文脈自由文法はよく使われる
・文脈自由文法では表現しきれない自然言語現象
（例）スイスドイツ語の従属節における交差依存

・複雑な RNA 塩基対の構造表現 (シュードノット)
・多重文脈自由文法は穏やかな文脈自由文法の拡張
・多項式時間解析可能
・各非終端記号は文字列の組を導出する

$B \rightarrow f(C,D)$ with $f(\langle x_1,x_2,x_3\rangle, \langle y\rangle) = \langle x_1 a y x_2, x_3\rangle$

ex. $\{ a^mb^nc^md^n \mid n,m > 0 \}$, $\{ ww \mid w \in \Sigma^* \}$

成果

	正　例		所属性質問と等価性質問	正例と所属性質問
文脈自由言語	可代入言語 (Clark & Eyraud'07)	k,l-可代入言語 (Yoshinaka'08)	合同性言語 (Clark'10)	より豊かな言語 (Clark et al.'08)
多重文脈自由言語	多次元可代入言語 (Yoshinaka'09)		合同性言語 (Yoshinaka & Clark'10)	より豊かな言語 (Yoshinaka'10)

多次元可代入多重文脈自由言語の学習

・p次元可代入性：

$\boldsymbol{x}\times\boldsymbol{u},\ \boldsymbol{x}\times\boldsymbol{v},\ \boldsymbol{y}\times\boldsymbol{u} \in L \Rightarrow \boldsymbol{y}\times\boldsymbol{v} \in L$

ただし，$\boldsymbol{u},\boldsymbol{v}$ は$m(\leq p)$個組文字列，$\boldsymbol{x},\boldsymbol{y}$ はm重文脈

・正例からの学習
＜アルゴリズム＞
・各正例 $w = u_0v_1u_1 \ldots v_mu_m$ の部分多重語 $\boldsymbol{v} = (v_1,\ldots,v_m)$ でラベルづけされた非終端記号 $[\boldsymbol{v}]$ をつくる
・各非終端記号 $[\boldsymbol{u}],[\boldsymbol{v}_1],\ldots,[\boldsymbol{v}_m]$ について
$\boldsymbol{u} = f(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_m)$ ならば次の規則を持つ
$[\boldsymbol{u}] \rightarrow f([\boldsymbol{v}_1],\ldots,[\boldsymbol{v}_m])$
・ $\boldsymbol{x}\times\boldsymbol{u},\ \boldsymbol{x}\times\boldsymbol{v}$ がともに正例ならば次の規則を持つ
$[\boldsymbol{u}] \rightarrow [\boldsymbol{v}]$
・w が正例ならば次の規則を持つ
$S \rightarrow [w]$

（例）
$L = \{a^m\#b^n\$c^m\#d^n \mid m,n \geq 0\}$
は次の正例から学習可能
$\{a\#\$c\#,\ a\#b\$c\#d,\ aa\#b\$cc\#d\}$

より複雑な語句分布情報の利用

・Concept Lattice (Clark '09)
・言語 L, 文脈有限集合 C, 文字列有限集合 W
・飽和対 $(C',W') \subseteq (C,W)$:
・ $(u,v) \in C'$ iff $(u,v)\times W \subseteq L$
・ $w \in W'$ iff $C' \times w \subseteq L$
・非終端記号を飽和対で特徴付ける
→　語句分布情報による学習

（例）
Dyck 言語： $L = \{ \varepsilon, ab, aabb, abab, aaabbb, \ldots \}$
$W = \{ \varepsilon, a, b, ab \}$, $C = \{ (\varepsilon,\varepsilon), (a,\varepsilon), (\varepsilon,b), (a,b) \}$

飽和対：
$\top = (W, \varnothing)$, $\bot = (\varnothing, C)$,
$S = (\ \{\varepsilon,ab\},\ \{(\varepsilon,\varepsilon), (a,b)\}\ )$,
$A = (\ \{a\},\ \{(\varepsilon,b)\}\ )$, $B = (\ \{b\},\ \{(a,\varepsilon)\}\ )$

可能な CNF 型規則（ただし $\top$ と $\bot$ は無視）
$S \rightarrow SS \mid AB \mid \varepsilon,\ A \rightarrow SA \mid AS \mid a,\ B \rightarrow SB \mid BS \mid b$

・可能な飽和対の数は指数的
・飽和対の集合は疎
・ZDDによる文法表現？　seqBDD in ZDD？