



HOKKAIDO UNIVERSITY

Title	語句の分布情報に基づいた多重文脈自由文法の学習
Author(s)	吉仲, 亮
Description	ERATO 湊離散構造処理系プロジェクトシンポジウム (第1回) : 第9回情報科学技術フォーラム(FIT2010) イベント企画セッション. 2010年9月8日 (水). 九州大学伊都キャンパス.
Relation	2010年度科学技術振興機構ERATO湊離散構造処理系プロジェクト講究録. p.379.
Issue Date	2011-06
Doc URL	https://hdl.handle.net/2115/48364
Type	conference presentation
File Information	07.FIT_yoshinaka.pdf



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
 $(u_0, u_1, \dots, u_m) \times (v_1, \dots, v_m) = u_0v_1u_1 \dots v_mu_m$

多次元可代入多重文脈自由言語の学習

- ・ p 次元可代入性:

$$x \times u, x \times v, y \times u \in L \Rightarrow y \times v \in L$$
 ただし, u, v は $m(\leq p)$ 個組文字列, x, y は m 重文脈

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

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

dat mer d'chind em Hans es huus lönd hälfe aasriiche
 私達が子供にハンスの家を塗るのを手伝わせしめる事

- ・ 複雑な RNA 塩基対の構造表現 (シュドノット)
- ・ 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. $\{a^m b^n c^m d^n \mid n, m > 0\}, \{ww \mid w \in \Sigma^*\}$

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)

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: $L = \{ \epsilon, ab, aabb, abab, aaabbb, \dots \}$
 $W = \{ \epsilon, a, b, ab \}, C = \{ (\epsilon, \epsilon), (a, \epsilon), (\epsilon, b), (a, b) \}$
 飽和対:
 $\top = (W, \emptyset), \perp = (\emptyset, C),$
 $S = (\{ \epsilon, ab \}, \{ (\epsilon, \epsilon), (a, b) \}),$
 $A = (\{ a \}, \{ (\epsilon, b) \}), B = (\{ b \}, \{ (a, \epsilon) \})$
 可能な CNF 型規則 (ただし \top と \perp は無視)
 $S \rightarrow SS \mid AB \mid \epsilon, A \rightarrow SA \mid AS \mid a, B \rightarrow SB \mid BS \mid b$

- ・ 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?