# 階層化 Pitman-Yor 過程を用いた文脈を考慮した確率文脈自由文法の推定

## ～ 分布学習の実データへの適用にむけて ～

D-42
柴田 千尋
東京工科大学

**分布学習**　文を文脈と部分文字列とにわけ，そこから形式文法を学習する枠組み

目的：分布学習の考え方にノンパラメトリックベイズを適用し，
文の集合のみから予測精度の高い文脈自由文法を学習したい．

## 1 確率モデル

文 w の導出

$$S \Rightarrow^{r_1} x_1 A_1 \alpha_1 \Rightarrow^{r_2} x_2 A_2 \alpha_2 \Rightarrow^{r_3} \cdots \Rightarrow^{r_m} x_m = w$$

中に現れる文脈 $(x, \alpha)$ を考慮に入れた階層化Pitman-Yor過程を定義：

(|x|,|α|)-context

$P_{A,(x,\alpha)}(A \to BC|(x,\alpha))$

↓ base measure

$P_{A,\pi(x,\alpha)}(A \to BC|\pi(x,\alpha))$

↓ base measure

⋮

↓ base measure

$P_{A,(\lambda,\lambda)}(A \to BC|(\lambda,\lambda))$

HPYP-PCFGs

$= P_A(A \to BC)$

↓ base measure

$P_{A,1}(A \to B*)P_{A,2}(A \to *C)$

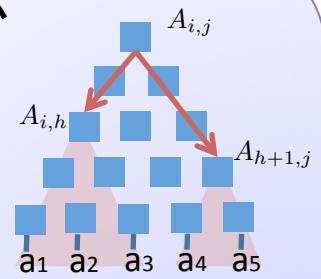
↓ base measure

$P_1(* \to B*)$　$P_2(* \to *C)$

↓ base measure

$\mathrm{Uniform}(V)$

## 2 サンプリング手法




blocked サンプリング

非終端記号が増えると極めて遅い．

ブロックあたりの計算量: $O(|V|^3|w|^3)$

提案するサンプリング手法

導出木の形(D)と，割り当てる非終端記号(A)にわけてサンプリング:

$$P(D|A,\tilde{\theta})P(A|\tilde{\theta})$$

ブロックあたりの計算量: $O(|V||w|^2 + |w|^3)$

## 3 予測精度での比較

適用データ： Brownコーパス

| methods | 予測精度 ( < - log P(w) > ) |
|---|---|
| Modified Kneser-Kney (4-gram) | 25.675 |
| HPYP-PCFGs with blocked sampler | 27.043 |
| (1,0)-context with proposed sampler | **25.596** |