2015年6月19日　統計数理研究所　オープンハウス

# Pitman-Yor隠れセミマルコフモデルによる教師なし完全形態素解析

**持橋大地**　　統計数理研究所 数理・推論研究系 学習推論グループ 准教授
daichi@ism.ac.jp

**内海慶, 塚原裕史**　デンソーITラボラトリ [共同研究]
{kuchiumi,htsukahara}@d-itlab.co.jp

## 1. 背景

Webには, 新語や口語, 新しい言い回しが無限に存在
→従来の教師あり学習で形態素解析するのは不可能

ローラのときに涙かブハァってなりました∩（´;ヮ;`）∩〜〜
真樹なんてこんな中２くさい事胸張って言えるぞぉ！
今日ね！らんらんとるいとコラボキャスするからおいで〜(*´∀`)ノシ
どうせ明日の昼ごろしれっと不在表入ってるんだろうなぁ。
テレ東はいつものネトウヨホルホルVTR鑑賞番組してんのか

Twitterにおけるテキストの例.

⇓

教師なし形態素解析 (持橋+ 2009). ただ品詞がない

- 単語には, 動詞／名詞／形容詞などの品詞が存在
- 単語分割には, 品詞を知ることが有用 (逆は明らか)
  例: “すももももももものうち”
- 未知の言語でも, その単語の種類を推定したい
  (計算言語学)

## 2. 提案法

文字列 $s$ が与えられたとき, それを分割した単語列 $\mathbf{w}$ および対応する品詞列$\mathbf{z}$ の確率を最大化:

$$\underset{\mathbf{w},\mathbf{z}}{\operatorname{argmax}}\ p(\mathbf{w},\mathbf{z}|s) \propto p(\mathbf{w},\mathbf{z},s) \qquad (1)$$

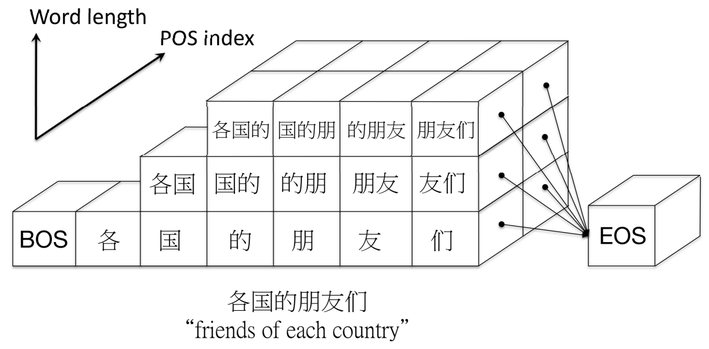
$p(s,\mathbf{w},\mathbf{z})$ の同時生成モデル

$z_0$=BOS; $s=\epsilon$ (an empty string).
**for** $t = 1 \cdots T$ **do**
  Draw $z_t \sim p(z_t|z_{t-1})$ ,
  Draw $w_t \sim p(w_t|w_1 \cdots w_{t-1}, z_t)$ ,
  Append $w_t$ to $s$ .
**end for**

隠れ状態を持つセミマルコフモデル(分割モデル)…
PYHSMM (Pitman-Yor Hidden semi-Markov Models).

## 3. モデルの学習

ベイズ学習: 文字列の集合が与えられた時, Forward-filtering Backward-sampling (Scott 2002) で動的計画法+MCMC. $\alpha[t][k][z]$ : 時刻$t$までの部分文字列 $c_1 \cdots c_t$ のうち, 最後の$k$文字が単語で, かつその品詞




各国的朋友们
“friends of each country”

が$z$である確率

$$\alpha[t][k][z] = \sum_{j=1}^{L}\sum_{y=1}^{K} p(c_{t-k}^{t}|c_{t-k-j+1}^{t-k}, z)\, p(z|y)\alpha[t-k][j][y]\,. \qquad (2)$$

計算:　Amazon　AWS のクラスタ計算機で並列
MCMC: 数日〜1週間程度. メモリ数10GB

## 4. 実験結果

日本語, 中国語, タイ語の教師なし単語分割において世界最高精度.

| Dataset | PYHSMM | NPY | BE | HMM$^2$ |
|---|---|---|---|---|
| Kyoto | **71.5** | 62.1 | 71.3 | NA |
| BCCWJ | **70.5** | NA | NA | NA |
| MSR | **82.9** | 80.2 | 78.2 | 81.7 |
| CITYU | **82.6**$^*$ | 82.4 | 78.7 | NA |
| PKU | **81.6** | NA | 80.8 | 81.1 |
| BEST | **82.1** | NA | **82.1** | NA |

NPY: Nested Pitman-Yor (Mochihashi+ 2009)
BE: Branching Entropy (Zhikov+ 2010)
HMM$^2$: Char and word HMMs (Chen+ 2014)

**例: 三河弁の解析**

三河弁(愛知)のTwitterのテキストから教師なし学習
(単語が何であるかは不明; 辞書を全く使わない)

| $z$ | Induced words |
|---|---|
| 2 | の 、 は に が で と も を 「 |
| 3 | ぞん かん ね のん だに だん りん か ん だのん |
| 9 | (*^^*) !(^-^; (^_^;) (^^;; !(^^;; |
| 10 | 。 ！ !! ？ 」 (≧∇≦) !!」「 |
| 11 | 楽 入 ど寒 大丈夫 会 受 停電 良 美味 台風が |
| 13 | に ら わ な よ ね だら じゃんね え ぁ |
| 41 | 豊橋 名古屋 三河 西三河 名古屋弁 名古屋人 大阪 |

## References

[1] Kei Uchiumi, Hiroshi Tsukahara, Daichi Mochihashi. “Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models”, ACL-IJCNLP 2015, *to appear*, 2015.
[2] 「隠れセミマルコフモデルに基づく品詞と単語の同時ベイズ学習」. 内海慶, 塚原裕史, 持橋大地. 情報処理学会自然言語処理研究会 NL-220, 2015.