統計的自然言語処理と機械学習

持橋大地　　数理・推論研究系 准教授　　daichi@ism.ac.jp

"*Linguista sum: linguistici nihil a me alienum puto.*", Roman Jakobson.
統計的自然言語処理を中心に, 主に離散データの統計モデルの研究をしています. 特に, 離散データの裏に潜む連続性をいかに扱うかや, 人間がカテゴリーをどうやって認識しているかに興味があります. 必然的にモデルは教師なし学習を基にしたものになりますが, 教師データとみなせる(自然な)データを統計モデルに取り入れることに躊躇はありません.
キーワード: ノンパラメトリックベイズ法, ガウス過程, ポアソン過程, MCMCなど

統計的自然言語処理

言語の生文字列のみから "単語" を学習する「教師なし形態素解析」(Mochihashi+, ACL 2009) の理論をさらに進め, 条件付確率場(CRF)との統合モデルにより, 少数の教師データを生かした半教師あり形態素解析の研究を行いました. ((株)博報堂との共同研究) (Fujii+, TACL 2017)

$$
\begin{aligned}
p(c_\ell^k|\mathbf{s}) &= \sum_j p(c_\ell^k, c_{\ell-j}^{\ell-1}|\mathbf{s}) \\
&\propto \sum_j \alpha[\ell-1][j]\cdot\beta[k][k-\ell+1]\cdot \\
&\qquad \exp\Bigl[\lambda_0 \log p(c_\ell^k|c_{\ell-j}^{\ell-1}) + \gamma[\ell,k+1)\Bigr] \\
&= \frac{\alpha[k][k-\ell+1]\cdot\beta[k][k-\ell+1]}{Z(\mathbf{s})}
\end{aligned}
$$

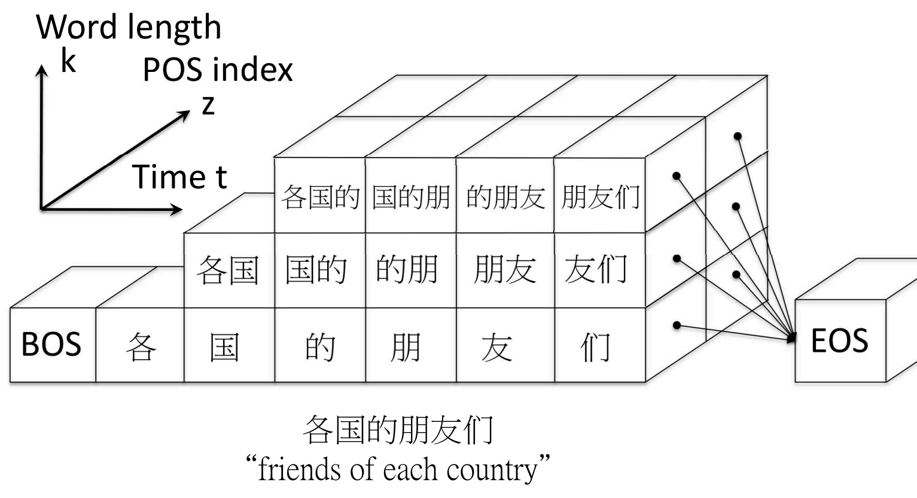
(左) ノンパラメトリックベイズ法との統合モデルによる前向き–後向き確率.

东软集团	19
游景玉	17
任尧森	17
南昆铁路	16
东方红三号"卫星	13
刘积仁	13
ｉｎｔｅｒｎｅｔ	11
东宝	11
张肇群	10
彭云	10
玲英	10
抚州	10
亚仿	10
南丁格尔	9
中远香港集团	7

にゃん	6
セフレ	6
フォロワー	5
https	4
December	4
トナカイ	3
アオサギ	3
フォロバ	3
じゅん	3
環奈	3
リプ	3
トッキュウジャー	2
リフォロー	2
酔っ払い	2
ツイート	2
クシャミ	2
エタフォ	2

(右) 中国語微博および日本語Twitterから学習された単語.

また, 単語への分割だけでなく, 各単語の品詞も教師なし学習する完全教師なし形態素解析の研究を行いました. (デンソーアイティラボラトリ(株)との共同研究) (ACL 2015)


Word length
k
POS index
z
Time t
各国的
国的朋
的朋友
朋友们
各国
国的
的朋
朋友
友们
BOS
各
国
的
朋
友
们
EOS
各国的朋友们
"friends of each country"

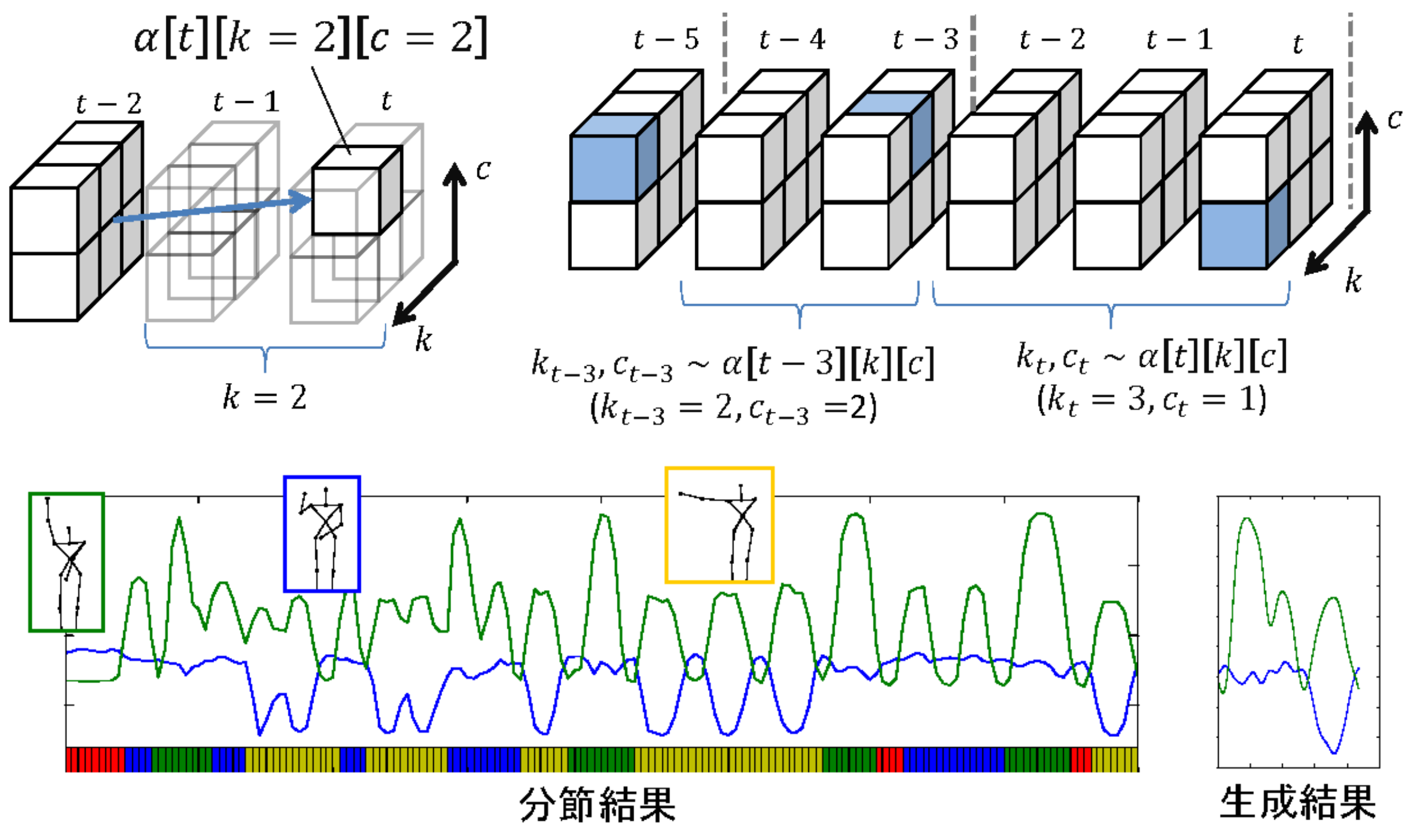
z	Induced words
2	の 、 は に が で と も を 「
3	ぞん かん ね のん だに だん りん か ん だのん
9	(*^^*) ！(^-^; (^_^;) (^^;; ！(^^;;
10	。 ！ !! ？ 」 (≧∇≦) !! 」「
13	に ら わ な よ ね だら じゃんね え ぁ
19	さん ん め 食べ って よろしく ありがとう
24	三河弁 この よ お 何 そ ほい 今日 また ほ
29	(！(; (´・ !! (*` ？(´・ (*^_^*)
32	((*\\(^ ＼(^ (^ ！*\\(^ 〜 (^_^ (*^
35	な ー そう 好き こと らん なん ら み 意味
36	いい ど うまい 杏果 ぐろ めっちゃ かわい はよ
41	豊橋 名古屋 三河 西三河 名古屋弁 名古屋人

Unsupervised analysis of a Japanese local dialect by PYHSMM. ($K=50$)

(上) 動的計画法による文字列の分節化. (右) 三河弁Twitterから教師なし学習された「単語」とその品詞.

ロボティクス

上の教師なし形態素解析を応用して, 各分節で文字列ではなく, ガウス過程による軌道を生成することで, ロボティクスにおいて関節の角度時系列ベクトルから「動作」を統計的に学習する研究を行っています. これは隠れセミマルコフモデルであるため各分節は潜在状態を持ち, さらに階層ディリクレ過程を用いることで潜在状態の数も自動的に推定可能としました. この一連の研究は, ロボティクスのトップ国際会議IROSにおいて本会議論文として発表されています (Nakamura+2017, Nagano+ IROS2018).

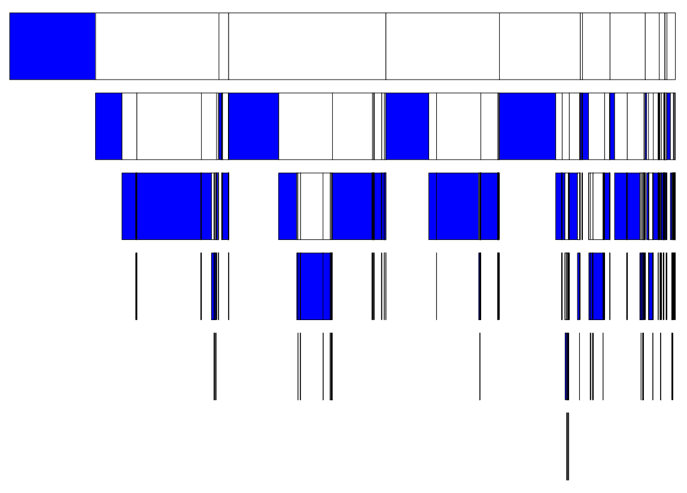

$\alpha[t][k=2][c=2]$
$t-2$
$t-1$
t
c
k
$k=2$
$t-5$
$t-4$
$t-3$
$t-2$
$t-1$
t
$k_{t-3}, c_{t-3} \sim \alpha[t-3][k][c]$
$(k_{t-3}=2, c_{t-3}=2)$
$k_t, c_t \sim \alpha[t][k][c]$
$(k_t=3, c_t=1)$
分節結果
生成結果

分節化の結果から得られた「動作」.

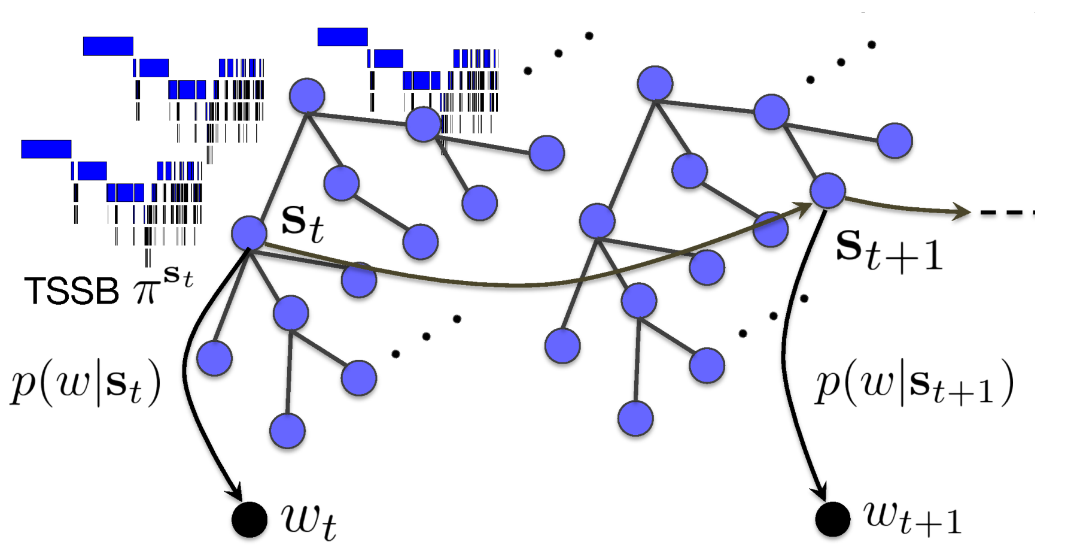
これは, 電通大・お茶の水女子大・阪大との共同研究です.

統計的機械学習

離散的な隠れマルコフモデル(HMM)は自然言語処理だけでなく, 経済学やバイオインフォマティクスなど様々な分野に使われる基本的な統計モデルですが, 状態が "2" "19" のように単純で, その間の相関が表せないという問題がありました. これに対し, 木構造Stick-breaking過程(Adams+2010)をさらに階層化することで, 無限の深さと無限個の分岐を持つ潜在的な木構造を状態空間に持つ無限木構造隠れマルコフモデルを定義し, 特別なMCMC法により学習が行えることを示しました. (持橋&能地2016) 本研究は2017年の情報処理学会山下研究賞を受賞しています.



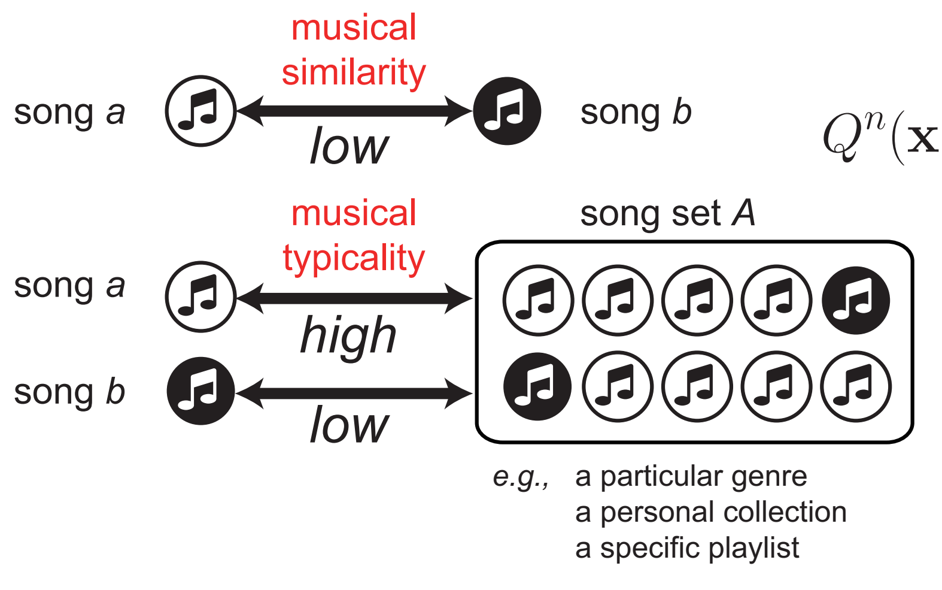
TSSB(1, 10, 0.2) からのサンプル.


$\mathbf{s}_t$
$\mathbf{s}_{t+1}$
TSSB $\pi^{\mathbf{s}_t}$
$p(w|\mathbf{s}_t)$
$p(w|\mathbf{s}_{t+1})$
w_t
w_{t+1}

無限木構造HMMの概要図.

音楽情報処理, 音声認識

音声認識においても, 音から教師データを介さず, 直接「単語」を認識する教師なし音声認識の研究を行っています (東工大・篠崎研究室との共同研究).
また, 音楽情報処理において曲がどれほど「ありがち」か, を定量化することは様々な意味で重要ですが, これを情報理論における典型集合としてとらえることで, 「ありがち度」を情報理論的尺度として計算する研究を行いました. この研究は, 音楽情報処理のトップ国際会議ISMIR 2016に採録されています (産総研 後藤グループとの共同研究).


song *a*
musical similarity
low
song *b*
song *a*
musical typicality
high
song *b*
low
song set *A*
e.g., a particular genre
a personal collection
a specific playlist

(a) 典型度と類似度の違い.

$$
\begin{aligned}
Q^n(\mathbf{x}) &= \prod_{i=1}^{n} Q(x_i) = \prod_x Q(x)^{N(x|\mathbf{x})} = \prod_x Q(x)^{nP(x)} \\
&= \prod_x \exp\bigl[nP(x)\log Q(x)\bigr] \\
&= \exp\Bigl[-n\Bigl(-\sum_x P(x)\log Q(x)\Bigr)\Bigr] \\
&= \exp\Bigl[-n\Bigl(H(P)+D(P||Q)\Bigr)\Bigr]. \quad \square
\end{aligned}
$$

(b) 典型集合による確率の計算.

その他の分野

この他にも, 教育工学, 計量政治学, 動物学, 言語学, 計量国語学, バイオインフォマティクスなどの分野で共同研究を行っています.

現在の学生：社会人学生2名 (NEC中央研究所, (株)AGC)
…いずれも博士後期課程
その他：日本学術振興会 学術情報分析センター研究員.