教師なし形態素解析とその周辺

持橋大地
統計数理研究所 数理・推論研究系
daichi@ism.ac.jp

ロボット学会データ工学ロボティクス研究会
2013-3-27(Wed)

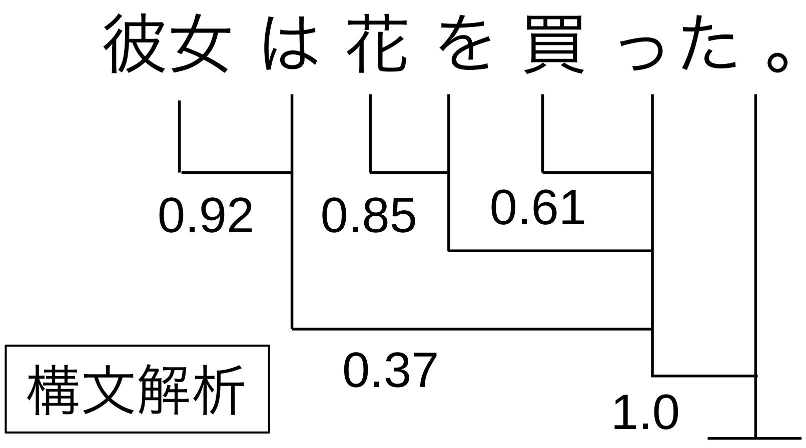
自己紹介

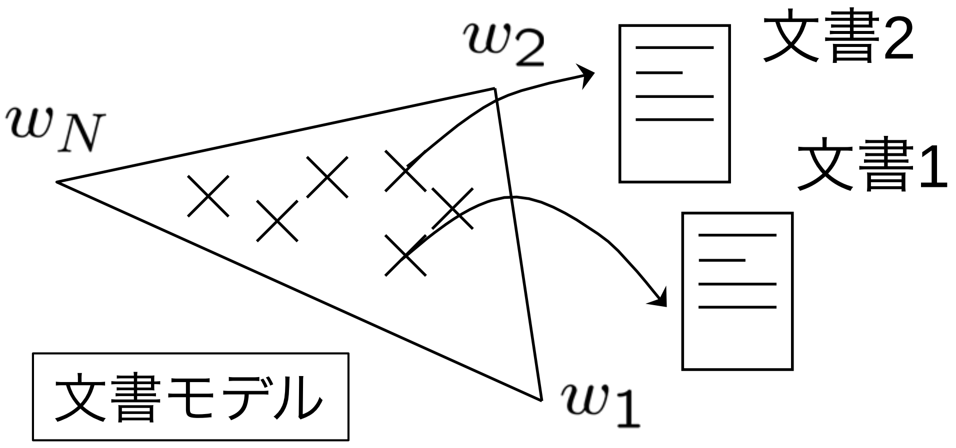
- 奈良先端大 (NAIST) 情報科学研究科 松本研
- ATR 音声研 研究員 (2003-2006)
- NTT CS研 研究員 (2007-2010)
- 統計数理研究所 (2011-)

専門：統計的自然言語処理

統計的自然言語処理

- 1990年代後半〜: 大量の言語データから、言語の性質を統計的に学習
  - 統計的機械学習の重要な一部
- 代表的な応用:
  構文解析、形態素解析、文書モデル、意味極性分類、照応解析、言語進化モデル、・・・・



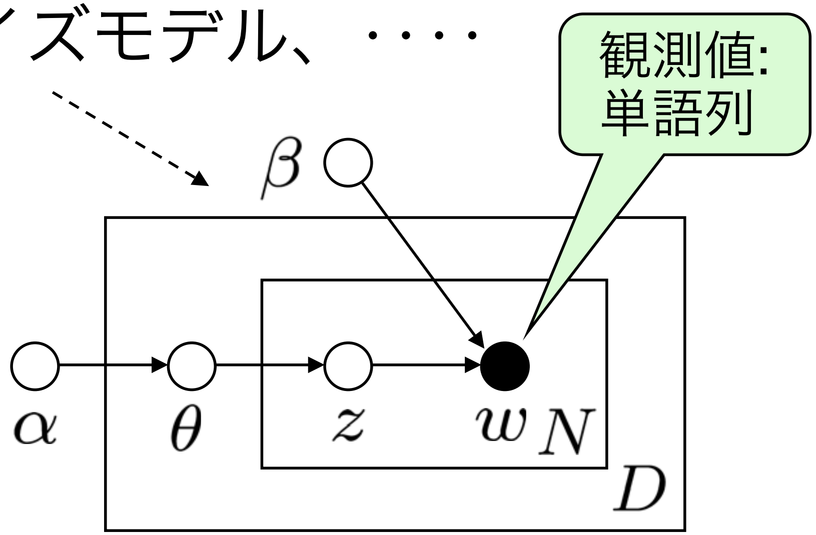


統計的自然言語処理でのモデル

- 論理式から、高度な統計モデルへ
  - チョムスキーの亡霊からの脱却
  - Webの登場と電子テキスト、計算資源の爆発的増大
  - 対数線形モデル、階層ベイズモデル、・・・・

$$p(t|\mathbf{x}, \mathbf{\Lambda}) = \frac{\exp\left(\sum_i \lambda_i f_i(\mathbf{x}, t)\right)}{\sum_{\mathbf{x}} \exp\left(\sum_i \lambda_i f_i(\mathbf{x}, t)\right)}$$

ある単語xの品詞が
形容詞である確率



- 統計的機械学習の一部として重要な地位を占める

統計的言語モデル

- 「自然言語の教師なし学習」の別名
- 単語列 $\mathbf{w} = w_1 w_2 \cdots w_N$ に対し、その確率 $p(\mathbf{w})$ を最大にする確率モデルを学習
  - nグラムとは限らない
  - 情報理論と密接な関係 (良いモデル≡良い符号化)
- 隠れ変数 $\mathbf{z}$ があってよい

$$p(\mathbf{w}) = \sum_{\mathbf{z}} p(\mathbf{w}, \mathbf{z})$$

  - $\mathbf{z}$ は何でもよい!!
  - 構文木、品詞列、感情極性、
    意味トピック、単語分割、etc,etc…

統計的言語モデル (2)

- 言語モデル研究の良い点：人手のラベルが一切不要
  - あらゆる言語データが、そのまま学習データになる
  - 新語、新表現、未知の言語、..のモデル化が可能
  - 統計的には一般に難しい
- Cons: 人手の正解と一致するとは限らない
  → 半教師あり学習
  - *これには、教師なしモデルが先に必要*

教師なし品詞解析

- 単語の持つ動詞, 形容詞, 名詞... などの品詞を同定することは、自然言語処理の多くの場面で有効

たなびく 雲 の 合間 から 漏れ出る 静かな 月 。

動詞 名詞 助詞 名詞 助詞 動詞 形容動詞 名詞

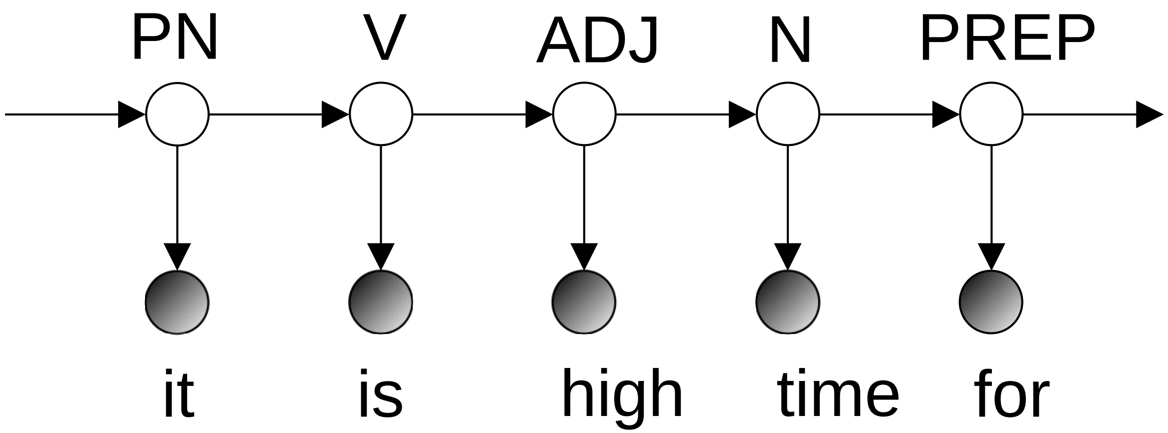
After procrastination, it is the time for prime-ministerial leadership .

PREP N PNV DT N PREP ADJ N

- 従来は, 上のような品詞タグを学習データに与えておき、識別器(＝回帰問題)を学習
- 問題: 上の「品詞」タグは充分に意味があるか?
  - 人手で大量のタグ付けをしなくとも, 自動的に「品詞」を学習できないか?

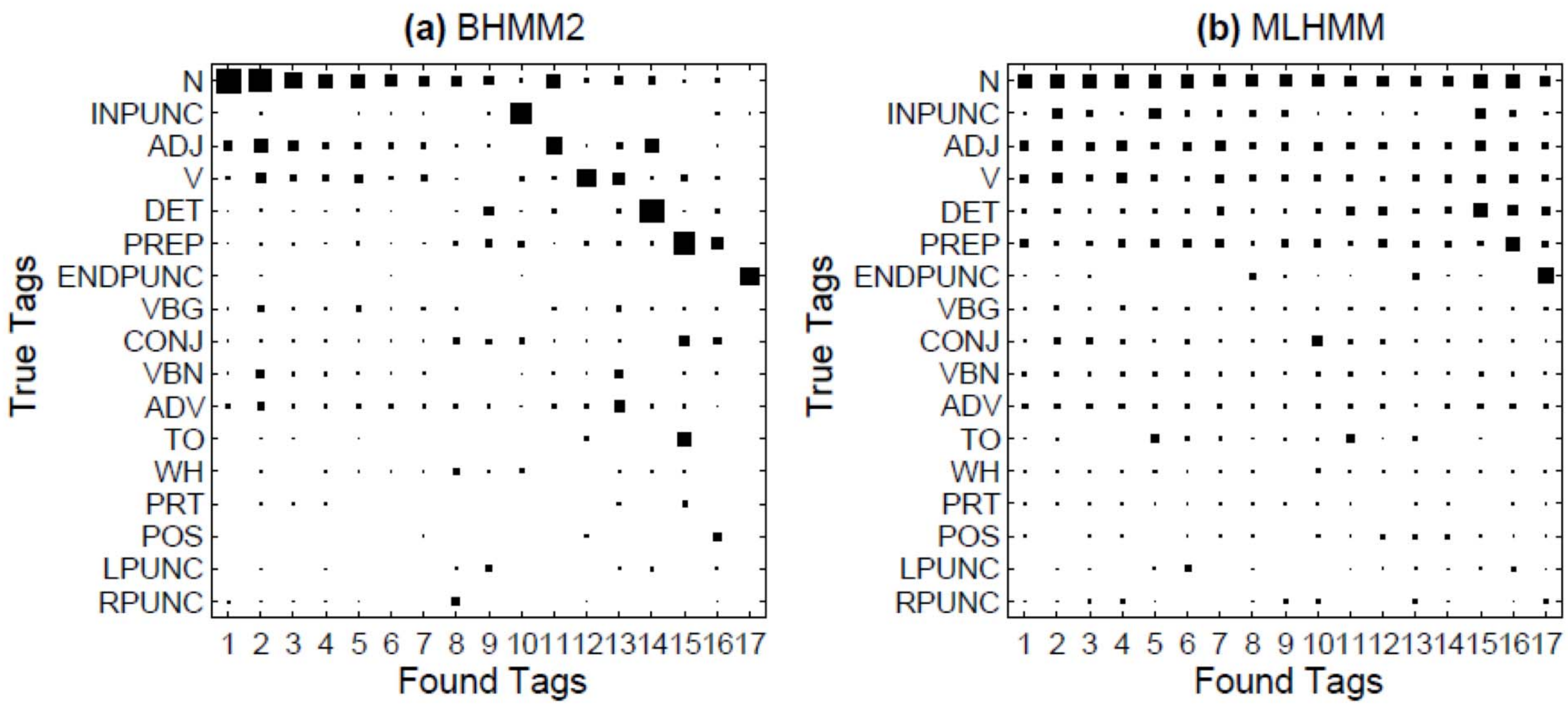
教師なし品詞解析 (2)

- 基本的なアプローチ: HMM (隠れマルコフモデル)を使おう (Johnson EMNLP 2007)
  - 「隠れ状態」が品詞に対応する(はず)



- 学習方法: EMアルゴリズム(Baum-Welch) or Gibbs
  - 1次Markovの場合は, 次の確率に従って状態yをサンプル

教師なし品詞解析 (3)

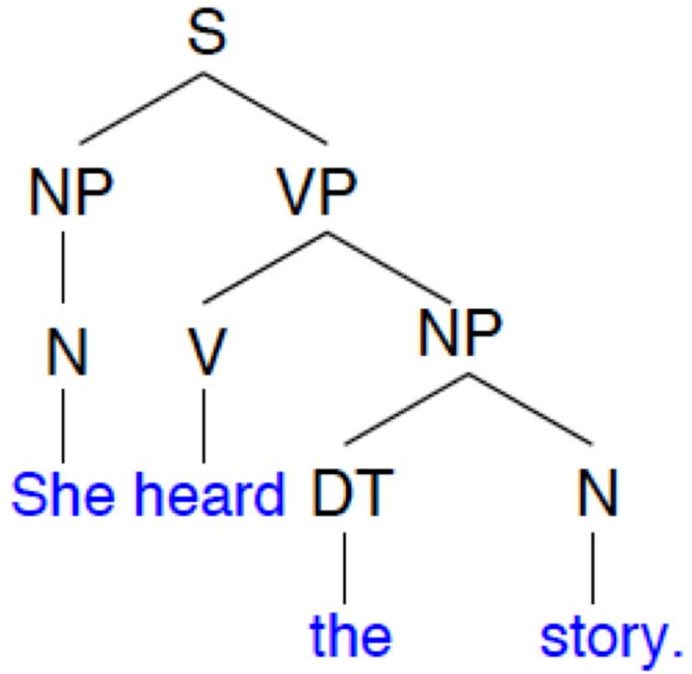


- 左側: ベイズHMM (Gibbs), 右側: EMアルゴリズム
- 最尤推定のEMアルゴリズムでは, 局所解に陥って良い解が見つかっていない (大データではEMも良い場合: EMNLP08)

教師なし構文解析 (1)

- 確率的文脈自由文法(PCFG)
  ・・・・与えられた文法の確率を学習し、確率最大となる
  　　構文木を解析

$$
\begin{array}{rcll}
S & \rightarrow & NP\ VP & : 0.9 \\
S & \rightarrow & VP & : 0.1 \\
VP & \rightarrow & V\ NP & : 0.75 \\
VP & \rightarrow & V & : 0.25 \\
NP & \rightarrow & DT\ N & : 0.8 \\
 & \cdots & &
\end{array}
$$



- 文法があれば、inside/outside確率テーブルとEMアルゴリズムで計算可能

教師なし構文解析 (2)

- 問題点: 文法は手で書いて与える必要がある
  - 自然言語をきちんと解析するためには、必要な文法は通常、非常に複雑になる

$$VP \rightarrow V^{motion}\ NP^{item} \quad : 0.5$$
$$VP \rightarrow V^{be}\ NP^{item} \quad : 0.174$$
$$VP \rightarrow V^{intrans}\ PP^{place} \quad : 0.22$$
$$\cdots$$

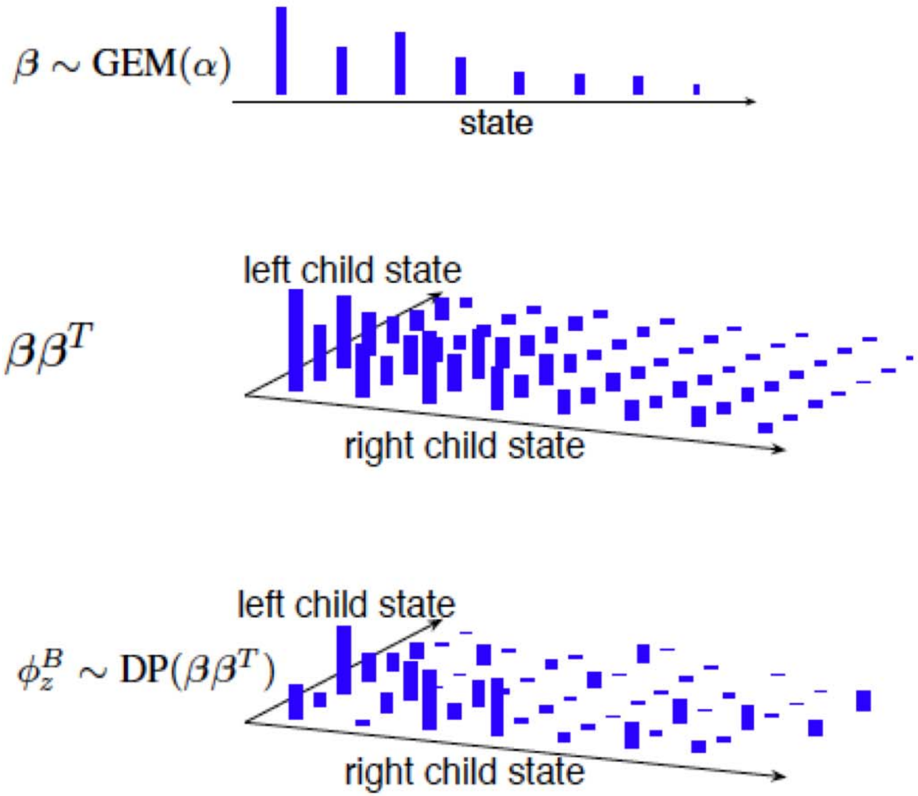
	V^{be}	V^{motion}	$VP^{intrans}$	NP^{item}	PP^{place}
V^{be}	0	0	0	0.174	0
V^{motion}	0	0	0	0.5	0
$VP^{intrans}$	0	0	0	0	0.22
NP^{item}	0.174	0.5	0	0	0
PP^{place}	0	0	0.22	0	0

- V^{motion},PP^{place}などのカテゴリはヒューリスティックに決めただけ (粗すぎ、細かすぎ)
  → 自動的に“カテゴリ”を決められないか?

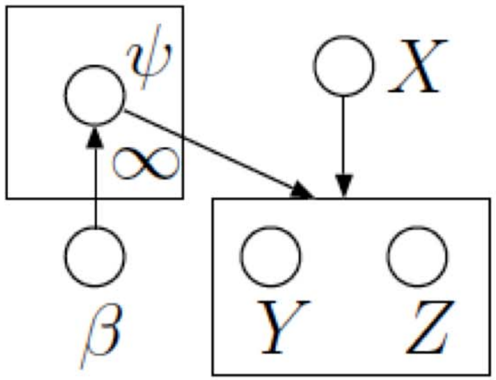
教師なし構文解析 (3)

学習は変分ベイズEM

- Liang+ (EMNLP 2007): HDPによる無限カテゴリPCFGの教師なし学習
  - 階層ディリクレ過程でカテゴリ生成＋状態遷移行列生成



$X \rightarrow Y\ Z$



F_1 measure = 87.08
with 428,375 rules,
EM F_1 = 88.36
with 706,157 rules

nグラムモデル

- nグラムモデル････古典的だが、音声認識や機械翻訳では未だ重要、基本的 (言葉のMarkovモデル)

$$p(\text{彼女 が 見る 夢}) = p(\text{彼女}) \cdot p(\text{が} \mid \text{彼女}) \cdot p(\text{見る} \mid \text{が}) \cdot p(\text{夢} \mid \text{見る})$$

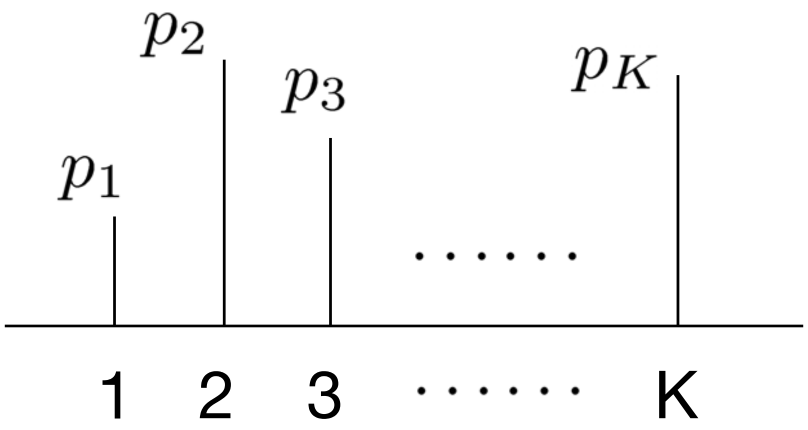
- nグラムモデルの問題: スムージング

現在のGoogle
カウント

$$\hat{p}(\text{yield}|\text{maximum likelihood will often}) = \frac{n(\text{maximum likelihood will often yield})}{n(\text{maximum likelihood will often})} = 0$$

  – 頻度そのままでなく、何か値を足したりする必要!

多項分布 (離散分布)

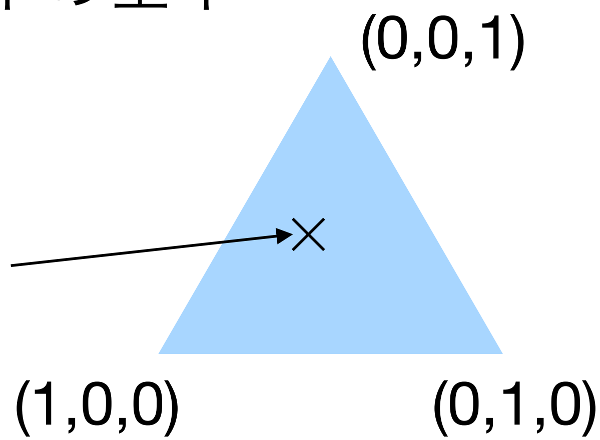


$$\sum_{k=1}^{K} p_k = 1,$$

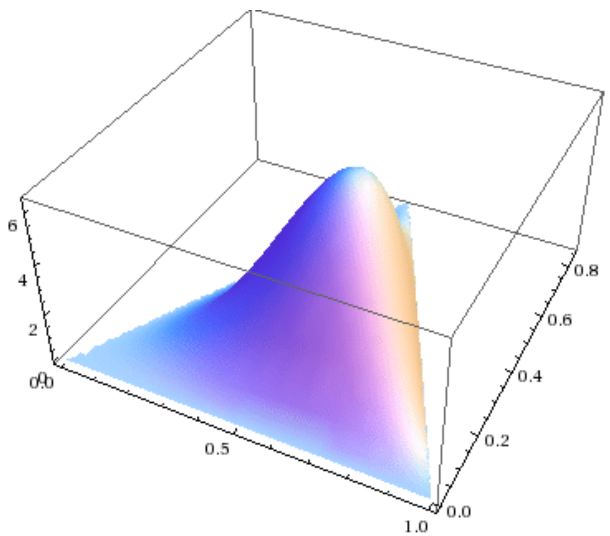
$$\forall k,\ p_k \geq 0$$

- K種類のアイテムのどれかが出る確率分布
  - 離散データの統計モデルの基本中の基本
- $\mathbf{p}$ は(K-1)次元の単体(Simplex)の内部に存在

$$\mathbf{p} = (p_1, p_2, \cdots, p_K)$$



ディリクレ分布



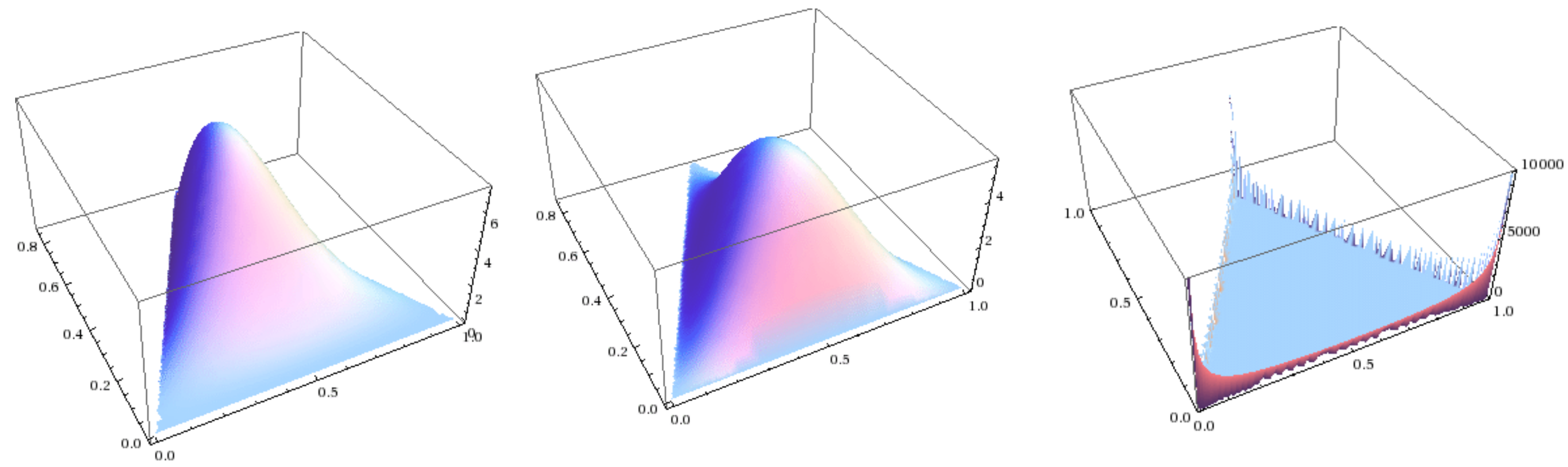
$$\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} p_k^{\alpha_k - 1}$$

パラメータ:

$$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$$

- ランダムな多項分布を生成する確率分布
- $\alpha_k \equiv 1$ のとき、単体上でUniformな分布
- 「期待値」：$\bar{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K) / \sum_k \alpha_k$
  「分散」　：$\alpha = \sum_k \alpha_k$

ディリクレ分布 (2)



$\boldsymbol{\alpha}=(2,4,2)$ $\boldsymbol{\alpha}=(2,2,2)$ $\boldsymbol{\alpha}=(0.5,0.5,0.5)$

- $\alpha_k > 1$ のとき、上に凸
- $\alpha_k < 1$ のとき、下に凸
  - 統計的自然言語処理等では、多くの場合 $\alpha \ll 1$ ($\alpha = 0.1 \sim 0.0001$ くらい)

ディリクレ分布に基づく予測

- ゆがんだ三面サイコロを振ったら、結果は $X=\{1,2,2,3,2,3\}$ (1=1回,2=3回,3=2回) だった。次の目は?
- ベイズの定理: $p(\mathbf{p}|X) \propto p(X|\mathbf{p})p(\mathbf{p})$

$$\propto \left(p_1^1 \cdot p_2^3 \cdot p_3^2\right) \cdot \left(\prod_{k=1}^3 p_k^{\alpha_k - 1}\right)$$

$$= p_1^{\alpha_1+1-1} \cdot p_2^{\alpha_2+3-1} \cdot p_3^{\alpha_3+2-1}$$

$$= \mathrm{Dir}(\alpha_1+1, \alpha_2+3, \alpha_3+2)$$

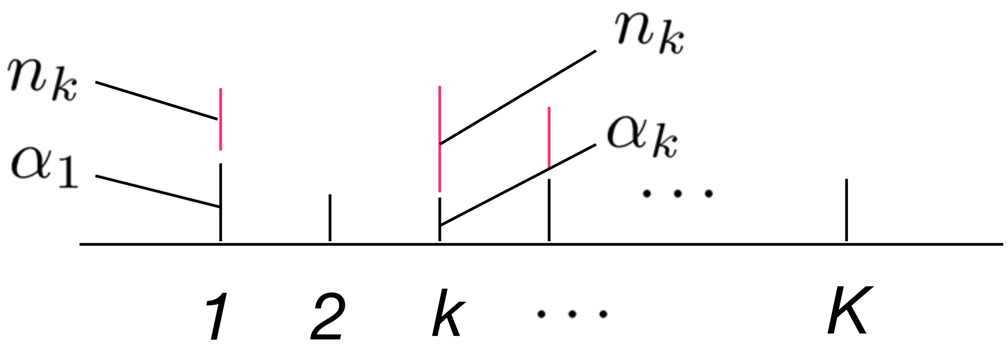
- $\mathbf{p}$ の期待値は、

$$E[\mathbf{p}|X] = \left(\frac{\alpha_1+1}{\alpha+6}, \frac{\alpha_2+3}{\alpha+6}, \frac{\alpha_3+2}{\alpha+6}\right) \quad (\alpha = \textstyle\sum_k \alpha_k)$$

ディリクレ分布に基づく予測 (2)

- 一般に、n回の観測の中でk番目のアイテムが n_k 回出現したとすると、

$$p(k|X) = \frac{\alpha_k + n_k}{\alpha + n}.$$



注: $n_k = 0$ のとき、$p(k|X) = \dfrac{\alpha_k}{\alpha + n}$

(出現しなかったアイテムにも正の確率)

ディリクレスムージング (MacKay 1994)

- nグラム確率分布 $p(w|h)$ にディリクレ事前分布 $p(\cdot|h) \sim \mathrm{Dir}(\boldsymbol{\alpha})$ を仮定すると、結果はα_wを足すのと同じ

$$p(w|h) = \frac{n(w|h) + \alpha_w}{\sum_w (n(w|h) + \alpha_w)}$$

  - $\boldsymbol{\alpha}$ はバイグラムなら、Newton法で最適化できる

- 問題: 性能が意外と低い
  - カウント n(w|h) が0のとき、

$$p(w|h) = \frac{\alpha_w}{\sum_w (n(w|h) + \alpha_w)}$$

  - $\alpha_w \ll 1$ なので、$1+\alpha_w \Longleftrightarrow \alpha_w$ に物凄い差!!

大体0.1〜0.001くらい

Kneser-Ney スムージング (Kneser,Ney 1995)

- 最高精度とよばれるスムージング法

$$p(w|h) = \frac{n(w|h) - D(n(w|h))}{n(h)} + \gamma(h) p(w|h')$$

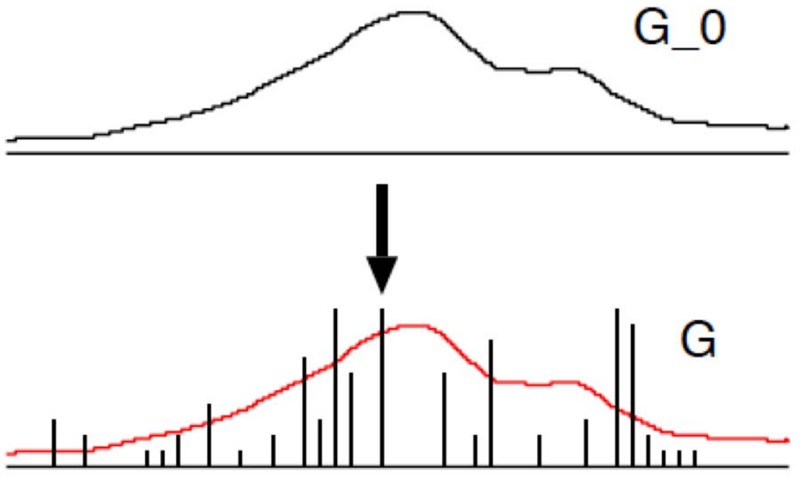
  - 頻度 n(w|h) から、一定数Dをディスカウント
  - $\gamma(h)$はhに後続する単語の種類数から決まる
- これは、下の階層Pitman-Yor過程による予測の近似であることが最近判明 (Goldwater+ 2006, Teh 2006)

$$p(w|h) = \frac{n(w|h) - d \cdot t_{hw}}{\theta + n(h)} + \frac{\theta + d \cdot t_h}{\theta + n(h)} p(w|h')$$

  - 階層Pitman-Yor過程とは?

ディリクレ過程 (Ferguson 1973)

- ディリクレ過程(Dirichlet process)とは、自然言語の1次元の場合、無限次元の多項分布を生成する分布のこと。

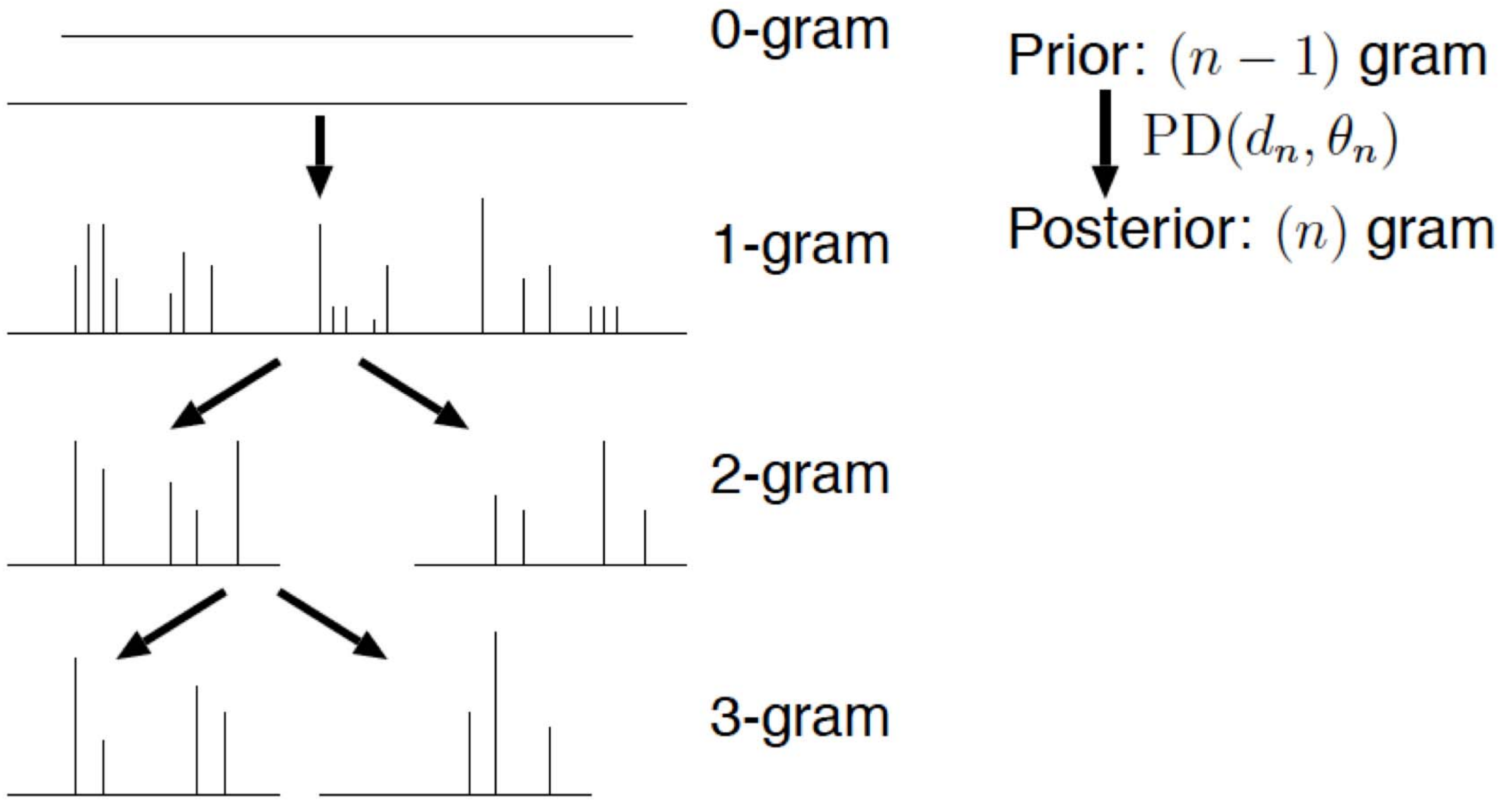


横軸:
可能な単語の種類

  - 元となる(連続)分布G_0に少し似た、無限次元の離散分布Gを生成
  - $G \sim \mathrm{DP}(\alpha, G_0)$ と表記 (α: 集中度パラメータ)
  - この2パラメータ拡張がPitman-Yor過程 $\mathrm{PY}(\alpha, d, G_0)$

階層Pitman-Yor過程

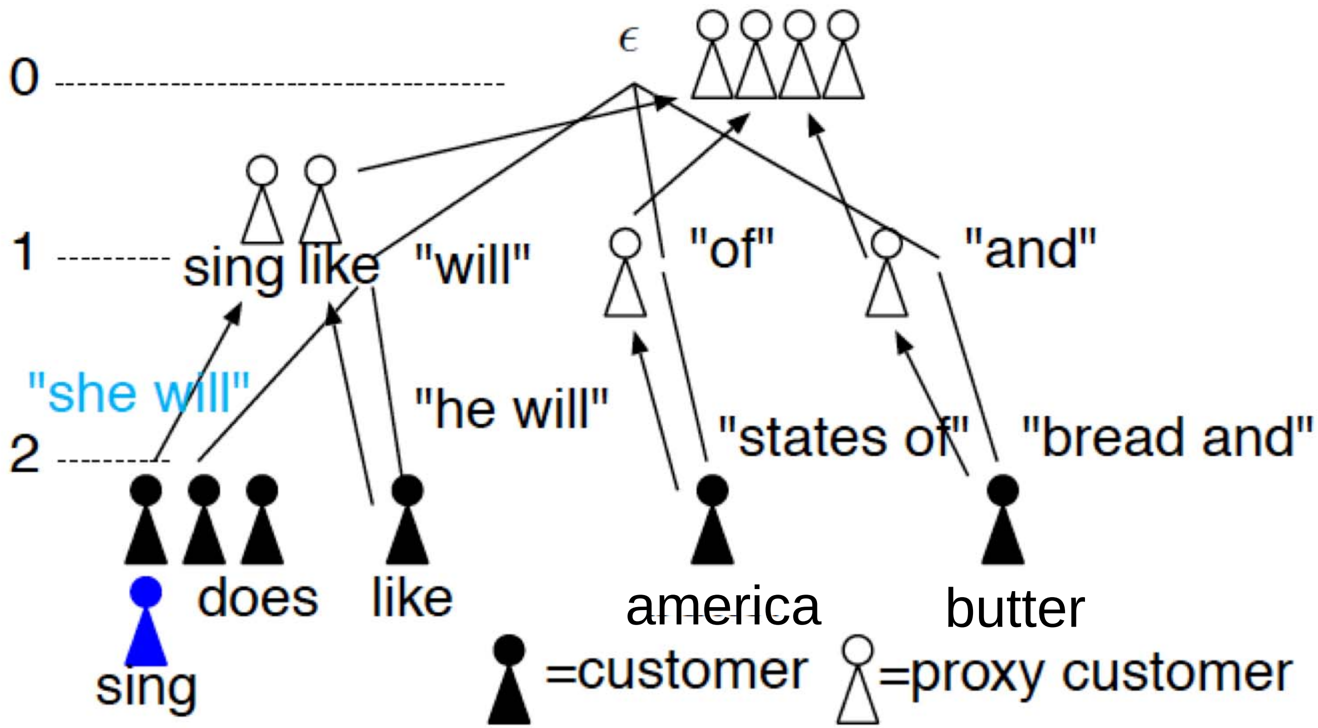
- nグラム分布が、階層的に(n-1)グラム分布からのPitman-Yor過程によって生成されたと仮定
  - 最初はUniform, だんだん急峻になる



階層CRP表現

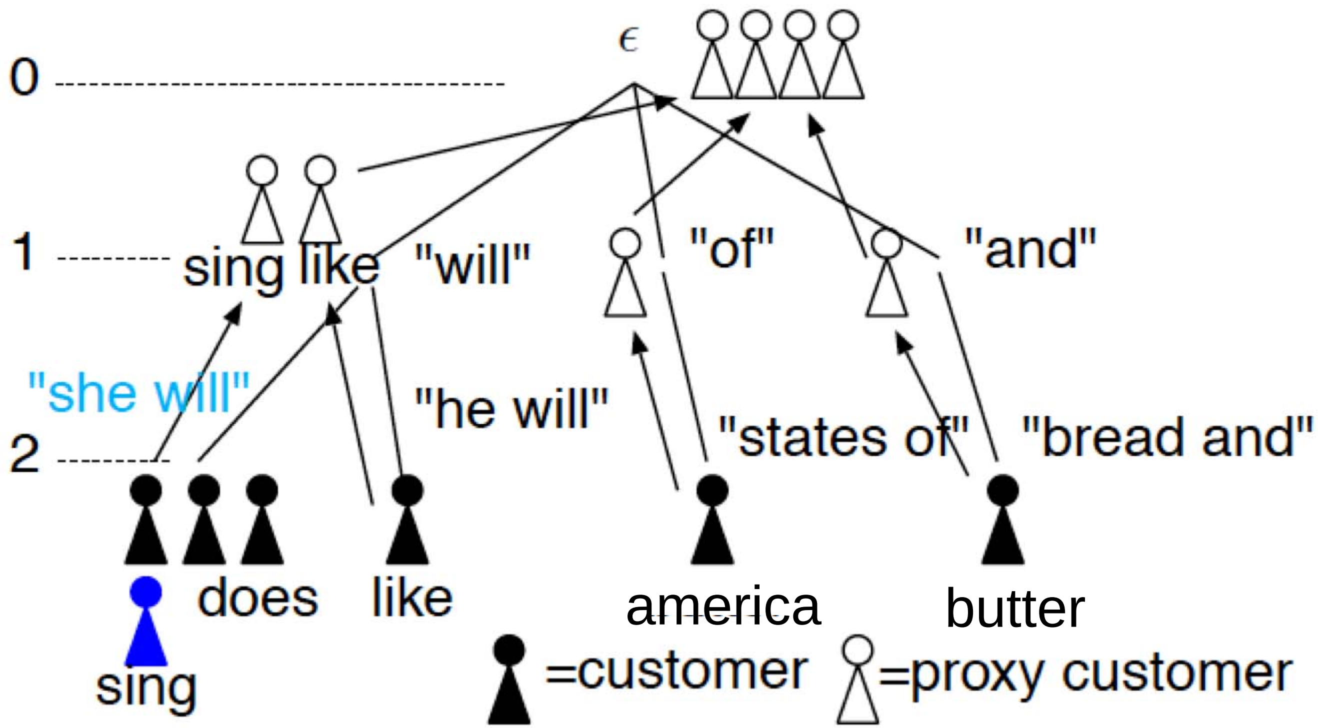
CRP=Chinese Restaurant Process

- 測度を直接扱う代わりに、カウントで離散表現する
  - 一人の「客」が1単語分のカウントに対応
  - 下の青い客は、文脈 “she will” の後に “sing” が1回現れたことを意味する (全部で2回)



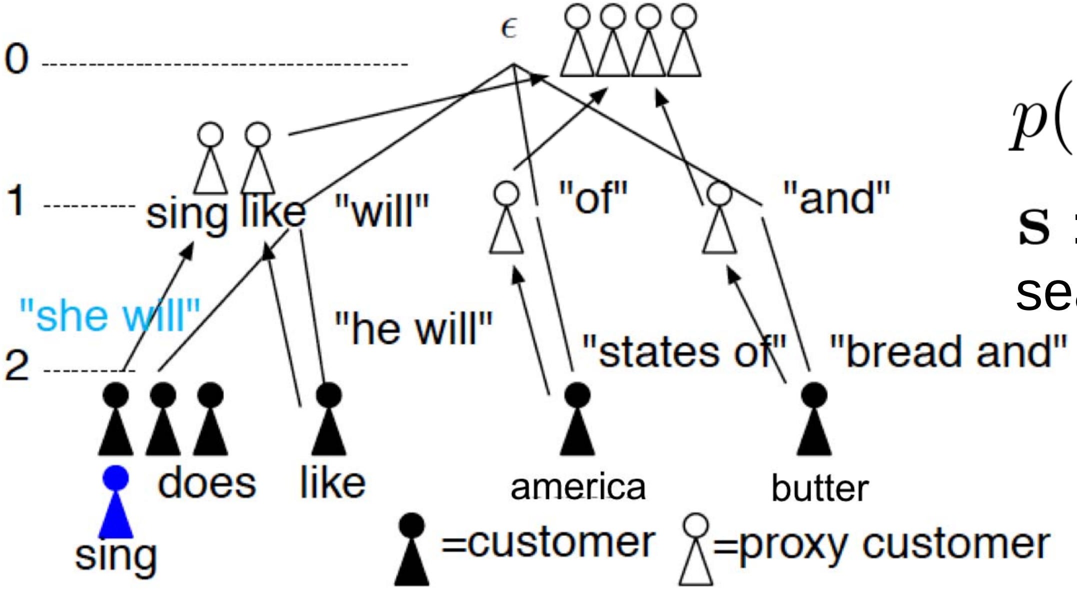
階層CRP表現 (2)

- 実際のカウント＝黒い客は、常に深さ(n-1)へ追加
  - 下のトライグラムの場合は、深さ2
- 適宜、確率的に、スムージングのためのカウントを親に再帰的に送る (白い客、代理客)



HPYLMの学習

- HPYLM (hierarchical Pitman-Yor language model) の学習＝潜在的な代理客の最適配置
- Gibbs sampling: 客を一人削除して再追加、を繰り返す
  - For each w = randperm(all counts in the corpus),
    - 客 w と関連する代理客をモデルから削除
    - 客 w をモデルに追加＝代理客を再サンプル

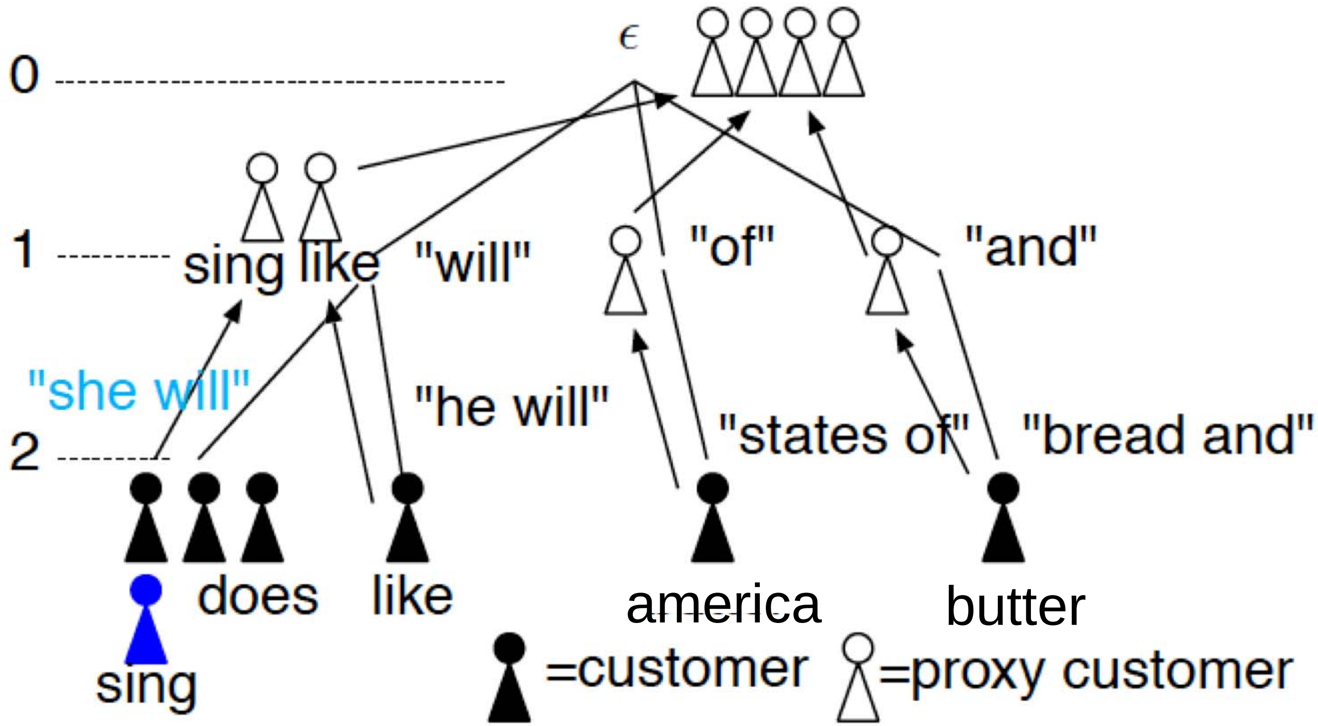


$$p(\mathbf{w}) = \sum_{\mathbf{s}} p(\mathbf{w}, \mathbf{s})$$

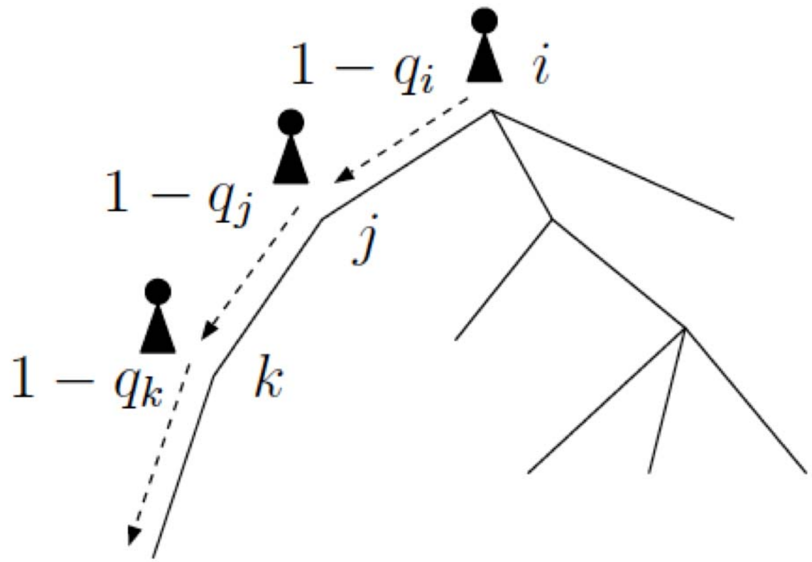
$\mathbf{s}$: 白い代理客の seating arrangements

HPYLM=nグラムモデルの問題

- 常に客を深さn-1に配置していいのか?
  - “other than”, “the united states of america” など、必要なnグラムのオーダーは本来異なるはず
  - HPYLMではどうすればいい?



VPYLM (Variable-order HPYLM)



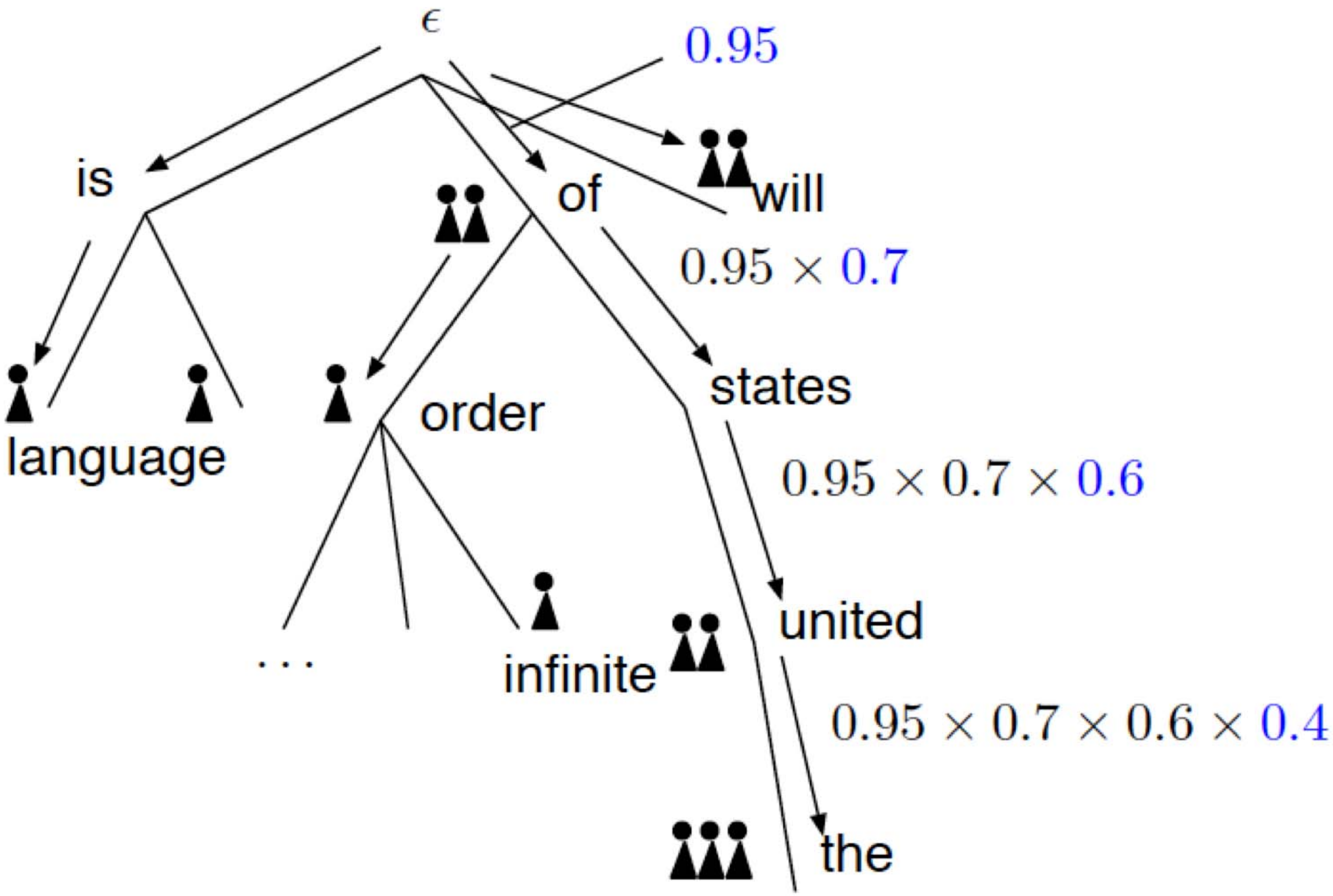
- 客を、木の根から確率的にたどって追加
- ノード i に,そこで止まる確率 q_iがある ($1-q_i$:通過確率)
  - q_i は、ランダムにベータ事前分布から生成

$$q_i \sim \mathrm{Be}(\alpha, \beta)$$

  - ゆえに、深さnで止まる確率は

$$p(n|h) = q_n \prod_{i=0}^{n-1} (1-q_i).$$

VPYLM, Variable-order HPYLM (2)



- “通過確率” (1-q_i)が大きい→深いノードに到達できる
- “通過確率” (1-q_i)が小さい→短いMarkov依存性を持つ

VPYLMの学習

- 学習データの各単語 $\mathbf{w} = w_1 w_2 \cdots w_N$ に, それを生んだ隠れたMarkovオーダー $\mathbf{n} = n_1 n_2 \cdots n_N$ が存在

$$p(\mathbf{w}) = \sum_{\mathbf{n}} \sum_{\mathbf{s}} p(\mathbf{w}, \mathbf{n}, \mathbf{s})$$

- Gibbs (MCMC)で $\mathbf{n}$ を推定

$$p(n_t | \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \propto \underbrace{p(w_t | n_t, \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{n_t\text{グラム予測確率}} \cdot \underbrace{p(n_t | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{\text{深さ}n_t\text{に到達するprior}}$$

  - 2つの項のトレードオフ (深いn_tにペナルティ)
  - 第二項の事前確率はどう計算する？

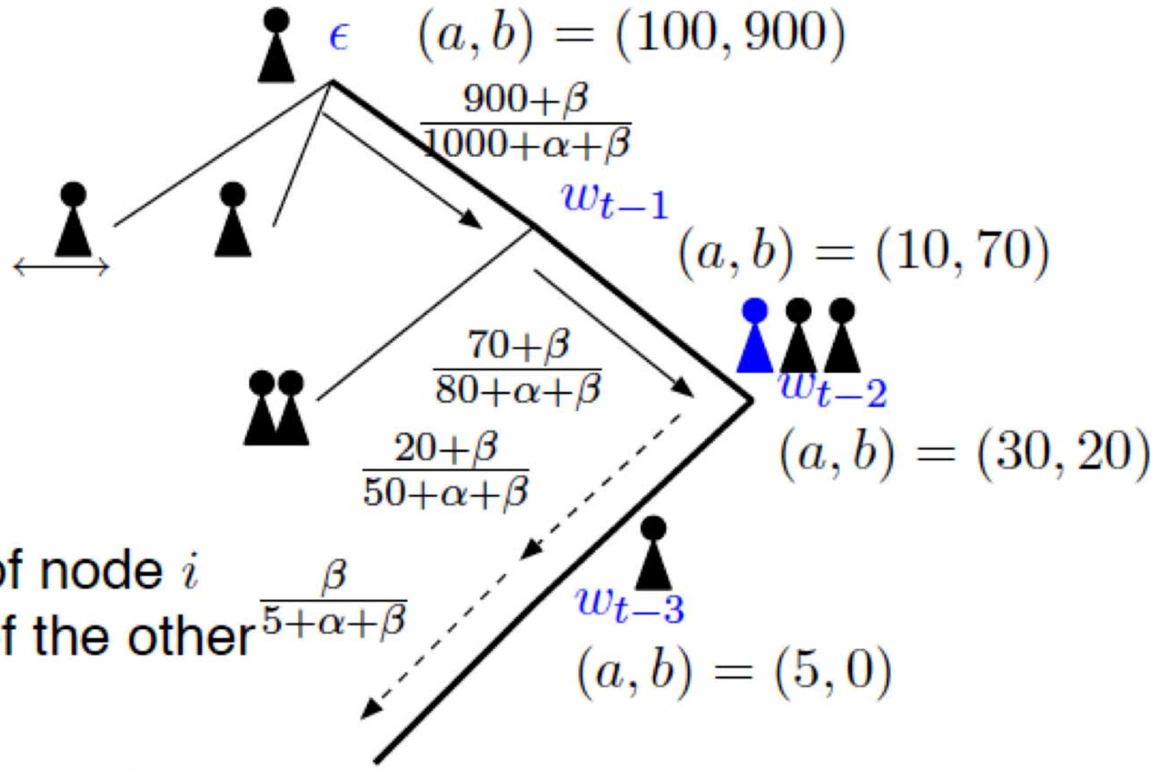
VPYLMの学習 (2)

w

$\cdots$	w_{t-2}	w_{t-1}	w_t	w_{t+1}	$\cdots$

n

$\cdots$	2	3	2	4	$\cdots$

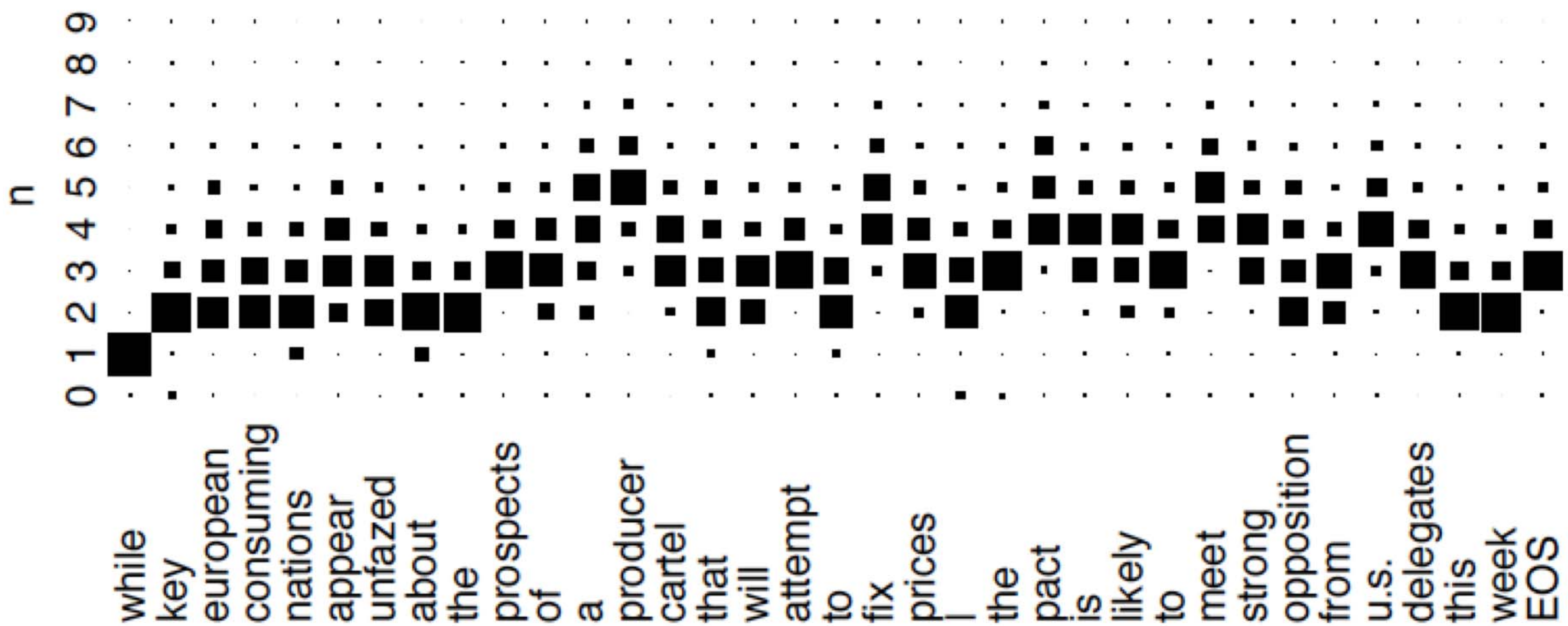


- We can estimate q_i of node i through the depths of the other customers
- Let a_i = # of times the node i was stopped at, b_i = # of times the node i was passed by:

$$p(n_t = n | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) = q_n \prod_{i=0}^{n-1} (1 - q_i) \tag{13}$$

$$= \frac{a_n + \alpha}{a_n + b_n + \alpha + \beta} \prod_{i=0}^{n-1} \frac{b_i + \beta}{a_i + b_i + \alpha + \beta}. \tag{14}$$

VPYLMの学習結果



- NAB (WSJ) コーパスの各単語が生成されたMarkovオーダーの推定結果
  - 情報量の多い語の後は短く、連語の後は長いなどの傾向が学習されている

VPYLMの予測

- 従来と異なり、nグラムオーダーnを事前に知らないので、nに関して積分消去

$$p(w|h) = \sum_n p(w,n|h) = \sum_n p(w|h,n)p(n|h)\,.$$

  - $p(n|h)$ は、先の計算で q_i から計算できる

- Suffix tree 上の Stick-breaking process になっている
  - 説明省略、NIPS 2011にほぼ同じアイデアがこの話を引かずに掲載

VPYLMの性能

n	SRILM	HPYLM	VPYLM	Nodes(H)	Nodes(V)
3	118.91	113.60	113.74	1,417K	1,344K
5	107.99	101.08	101.69	12,699K	7,466K
7	107.24	N/A	100.68	*27,193K*	10,182K
8	107.21	N/A	100.58	*34,459K*	10,434K
∞	—	—	100.36	—	10,629K

- SRILM: SRI言語モデルツールキット (Kneser-Ney)
- 少ないノード数で、高い性能
  - パープレキシティ＝平均予測確率の逆数
    (smaller is better)
  - ∞-gram が可能!! (今や、nは不要)

VPYLMからの生成

> ‘how queershaped little children drawling-desks, which would get through that dormouse!’ said alice; ‘let us all for anything the secondly, but it to have and another question, but i shalled out, ‘you are old,’ said the you’re trying to far out to sea.

(a) Random walk generation from a character model.

Character	s a i d ␣ a l i c e ; ␣ ‘ l e t ␣ u s ␣ a l l ␣ f o r ␣ a n y t h i n g ␣ ···
Markov order	5 6 5 4 7 10 6 5 4 3 7 1 4 8 2 4 4 6 5 5 4 4 5 5 6 4 5 6 7 7 7 5 3 3 4 5 9 ···

(b) Markov orders used to generate each character above.

- 「不思議の国のアリス」の∞-gram文字モデルからのランダムウォーク生成
  - 生成では、気をつけないと元データがそのまま再生されてしまう

「確率的フレーズ」の計算 (英語／WSJ)

p	Stochastic phrases in the suffix tree
0.9784	primary new issues
0.9726	^ at the same time
0.9556	american telephone &
0.9512	is a unit of
0.9394	to # % from # %
0.8896	in a number of
0.8831	in new york stock exchange composite trading
0.8696	a merrill lynch & co.
0.7566	mechanism of the european monetary
0.7134	increase as a result of
0.6617	tiffany & co.
:	

「確率的フレーズ」の計算 (日本語／毎日新聞)

p	Stochastic phrases in the suffix tree
0.9725	早けれ ば
0.9678	日 発表 し た
0.9592	来 年 # 月
0.9584	官房 審議 官
0.9185	いる に も かかわら
0.8877	震度 は 次 の 通り
0.8676	フェルメール と その 時代
0.8068	ある と し て いる 。
0.7870	こと が # 日 明らか に なっ
0.7661	記録 # 分 # 秒 #
0.7242	全豪オープン 第 # 日 は # 日
0.6472	^ 国際 オリンピック 委員 会 IOC
0.5019	ロッテ # 勝 # 敗 # 分 。
0.5000	行う こと を 明らか に し た
:	

最近の話

- 潜在的なMarkovオーダーを学習せず、データを「丸覚え」して、ベイズで最適な予測を行う "Sequence Memoizer" が提案 (Wood+, ICML 2009)
  - 途中のノードでの確率を復元する統計理論
- ただし、言語の性質は何も学習しない
  - 生成に用いると、常に最長の文脈を用いるため、学習データがそのまま出てくる
  - 学習データを全部覚えるのは無駄なのではないか?
  - 統計的には、非常に興味深い話

形態素解析

- 日本語や中国語等は単語に分けられていない
  ・・・・自然言語処理の非常に重要な課題

```
% echo “やあこんにちは, 大学内はどうですか。“
 | mecab -O wakati
やあ こんにちは , 大学 内 は どう です か 。
 × (やあこん にち は , 大 学内 はどう で す か 。)
```

  – Chasen, MeCab (NAIST)などが有名なツール
- これまで、教師あり学習 (supervised learning)によって学習されてきた
  – 人手で、単語分割の「正解例」を何万文も作成
  – 膨大な人手と手間のかかるデータ作成

形態素解析 (2)

```
# S-ID:950117245-006 KNP:99/12/27
* 0 5D
一方 いっぽう * 接続詞 * * *
、 、 * 特殊 読点 * *
* 1 5D
震度 しんど * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 2 3D
揺れ ゆれ * 名詞 普通名詞 * *
の の * 助詞 接続助詞 * *
* 3 4D
強弱 きょうじゃく * 名詞 普通名詞 * *
```

毎日新聞1995年度記事から38,400文(京大コーパス)の例

- 膨大な人手で作成した教師(正解)データ
  - 対数線形モデルやその拡張を用いて識別器を学習
- 話し言葉の「正解」？古文？未知の言語?
  - |女御|更衣|あ|また|さ|ぶら|ひ|た|ま|ひける|中|に|、|…

中国語の場合: 新浪微博 (Twitterクローン)

対@mo小颖 说：星期五吃饭~~！6点啊 [乐乐]
快下班了。。。。。可是我的时间才刚刚开始！
注册口碑卡~\(≧▽≦)/~啦啦啦
亲爱的，现在的我很成熟，我知道你对我意味着什么，我要好好珍惜你，我不要在错过一个值得我爱一生的女孩。
快艺考了。中西上戏，最差也要山艺锕。
哈哈哈……天真不嘛～可耐不嘛！雕塑课……明显是摆拍的-.-

- 中国語··日本語と同様に、まず単語に分ける必要
  - 膨大で無限に生まれる「未知語」！
- 新浪微博のユーザーは1億4千万人(!)

教師なし形態素解析

- 確率モデルに基づくアプローチ: 文字列 s について、それを分割した単語列 $p(\mathbf{w}|s)$ の確率

$$\hat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{w}|s)$$

  を最大にする $\hat{\mathbf{w}}$ を探す
  - 例: p(今日 は もう 見た) > p(今日 はも う見 た)
  - 教師データを使わない；辞書を使わない
  - 「言語として最も自然な分割」を学習する
- あらゆる単語分割の可能性を考える
  - たった50文字の文でも、2^50=1,125,899,906,842,624 通りの天文学的組み合わせ(さらに無数の文が存在)

文の確率: nグラムモデル

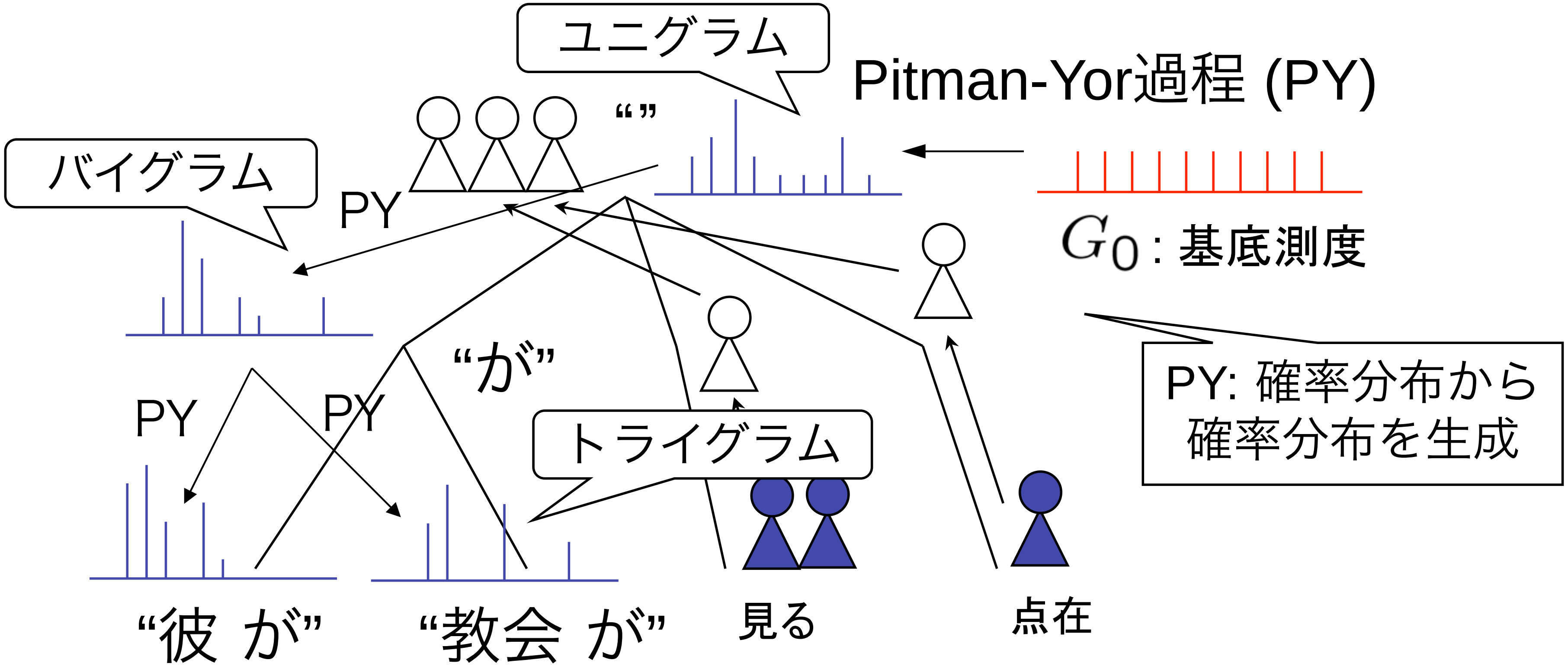
p(今日 は もう 見た)
= p(今日|^)・p(は|今日)・p(もう|は)・p(見た|もう)

文頭を表す特殊文字

- 条件付き確率の積で文の確率を計算
  - 自然言語処理では、きわめて強力 (Shannon 1948)
- 確率のテーブルは、ほとんどが０
  - 階層的なスムージングが不可欠
  - あらゆる部分文字列が「単語」になりうる

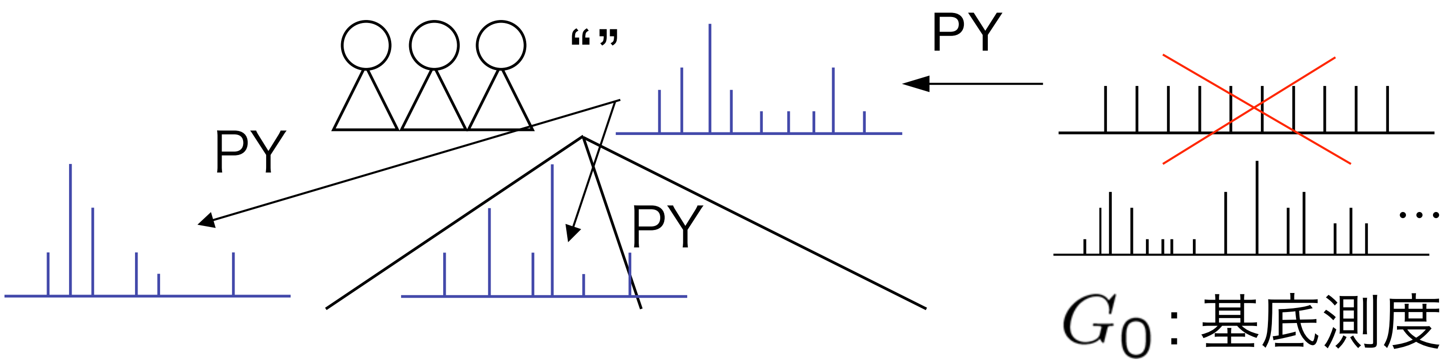
➡ 階層ベイズモデル: 階層Pitman-Yor過程言語モデル (HPYLM) (Teh 2006; Goldwater+ 2005)
• Pitman-Yor過程: ディリクレ過程 (GEM分布) の一般化

準備: HPYLM n-gram



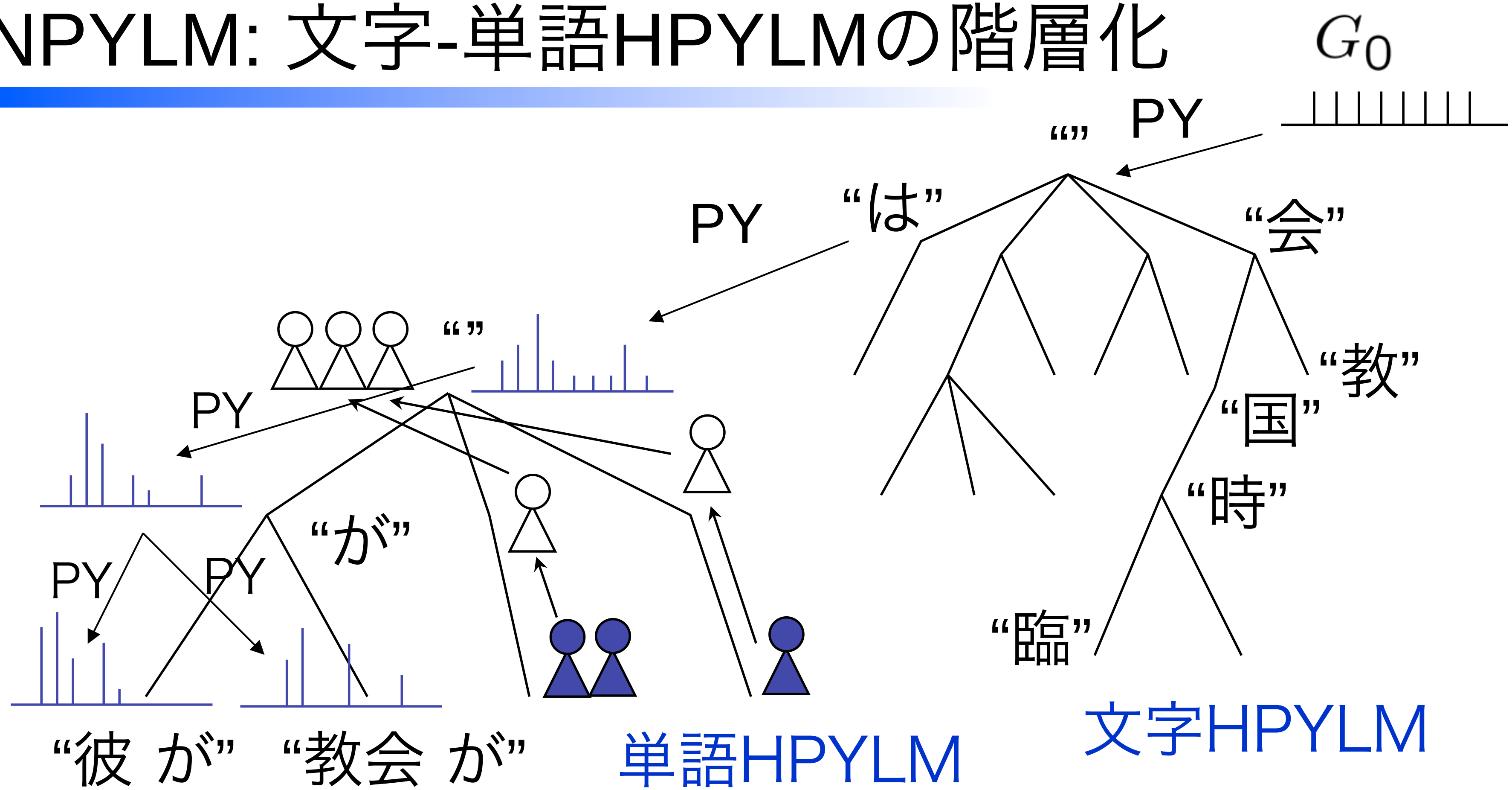
- カウントが0でも、より低いオーダーのMarkovモデルを用いて階層ベイズでスムージング
  - 注目している単語がそもそも存在しなかったら?

HPYLM: 無限語彙モデル



- 基底測度 G_0 は、単語の事前確率を表す
  - 語彙Vが有限なら、$G_0(w \in V) = 1/|V|$
- G_0 は可算無限でもよい！→ 無限語彙
  - PYに従って、必要に応じて「単語」が生成される
  - 「単語」の確率は、文字n-gram=もう一つのHPYLM
    - 他の方法で与えてもよい (が、再学習が面倒)

NPYLM: 文字-単語HPYLMの階層化



- HPYLM-HPYLMの埋め込み言語モデル
  - つまり、階層Markovモデル
- 文字HPYLMの G_0 は, 文字数分の1 (日本語なら1/6879)

NPYLMの学習問題の定式化

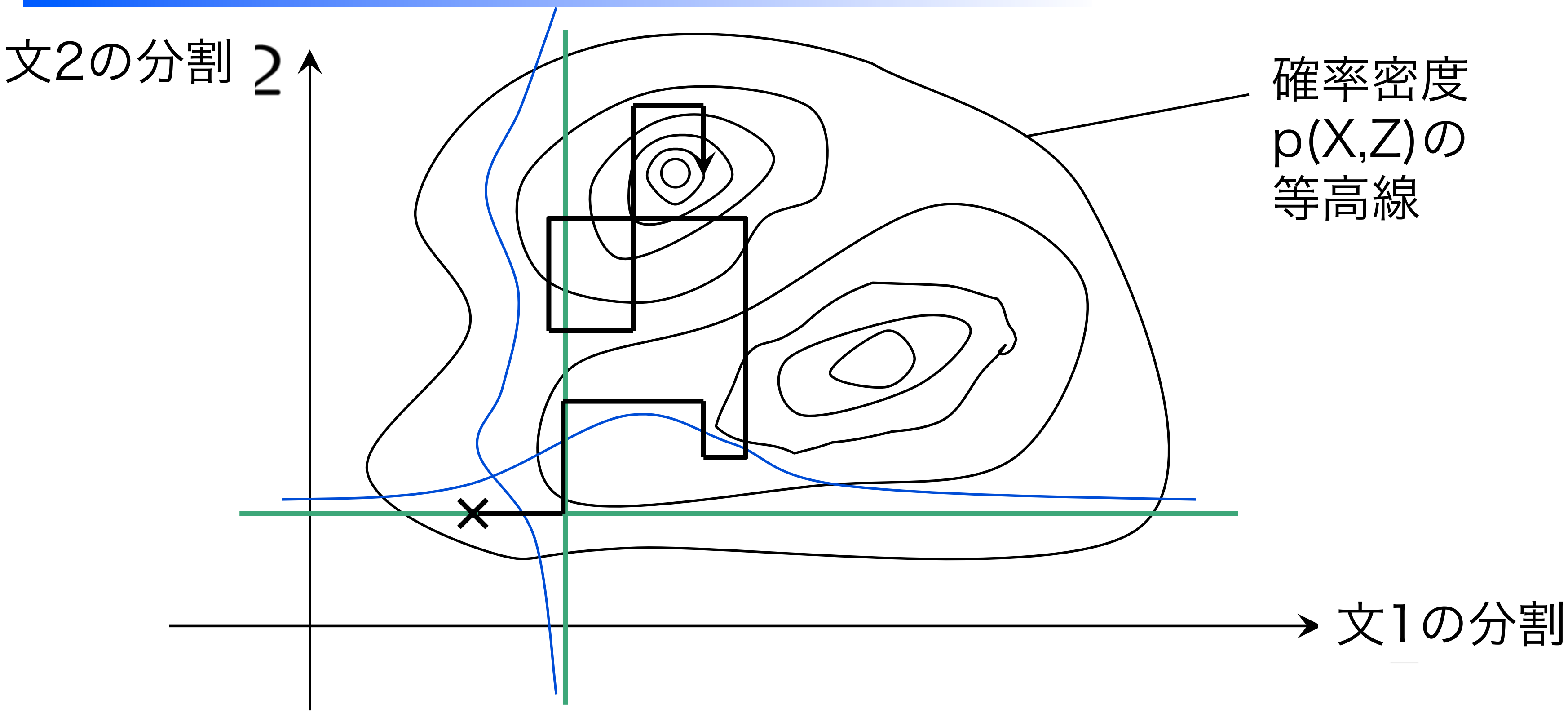
- データ: $\mathbf{X} = \{s_1, s_2, \cdots, s_X\}$ (文の集合)
  - 文: $s = c_1 c_2 \cdots c_N$ (文字列)
  - 隠れ変数: $\mathbf{z} = z_1 z_2 \cdots z_N$ ($z_i = 1$ のとき単語境界)
    - 隠れ変数の組み合わせは指数的に爆発
- 文がそれぞれ独立だと仮定すると、

$$p(\mathbf{X}) = \prod_{n=1}^{X} p(s_n) \qquad (1)$$

$$p(s_n) = \sum_{\mathbf{z}_n} p(s_n, \mathbf{z}_n) \qquad (2)$$

  - 各文s_nの分割 $\mathbf{z}_n$ を、どうやって推定するか？
    → ブロック化ギブスサンプリング、MCMC.

Blocked Gibbs Sampling



- 確率 p(X,Z) を最大にする単語分割を求める
- 単語境界は、前後の「単語」に強い依存関係
  → 文ごとに、可能な単語分割をまとめてサンプル
  (Blocked Gibbs sampler)

Blocked Gibbs Sampler for NPYLM

- 各文の単語分割を確率的にサンプリング
  →言語モデル更新
  →別の文をサンプリング
  ...を繰り返す.
- アルゴリズム:
  0. For s=s_1…s_X do
     parse_trivial(s,Θ).
  1. For j = 1..M do
     For s=*randperm*(s_1…s_X) do
       言語モデルからwords(s)を削除
       words(s) ～ p(w|s,Θ) をサンプリング
       言語モデルにwords(s)を追加して更新
     done.

文字列全体が一つの「単語」

Θ:言語モデルのパラメータ

Gibbs Samplingと単語分割

1 神戸では異人館　街の 二十棟 が破損した 。

2 神戸 では 異人館 街の 二十棟 が破損した 。

10 神戸 では 異人館 街の 二十棟 が破損した 。

50 神戸 で は異人 館 街 の 二 十 棟 が 破損 し た 。

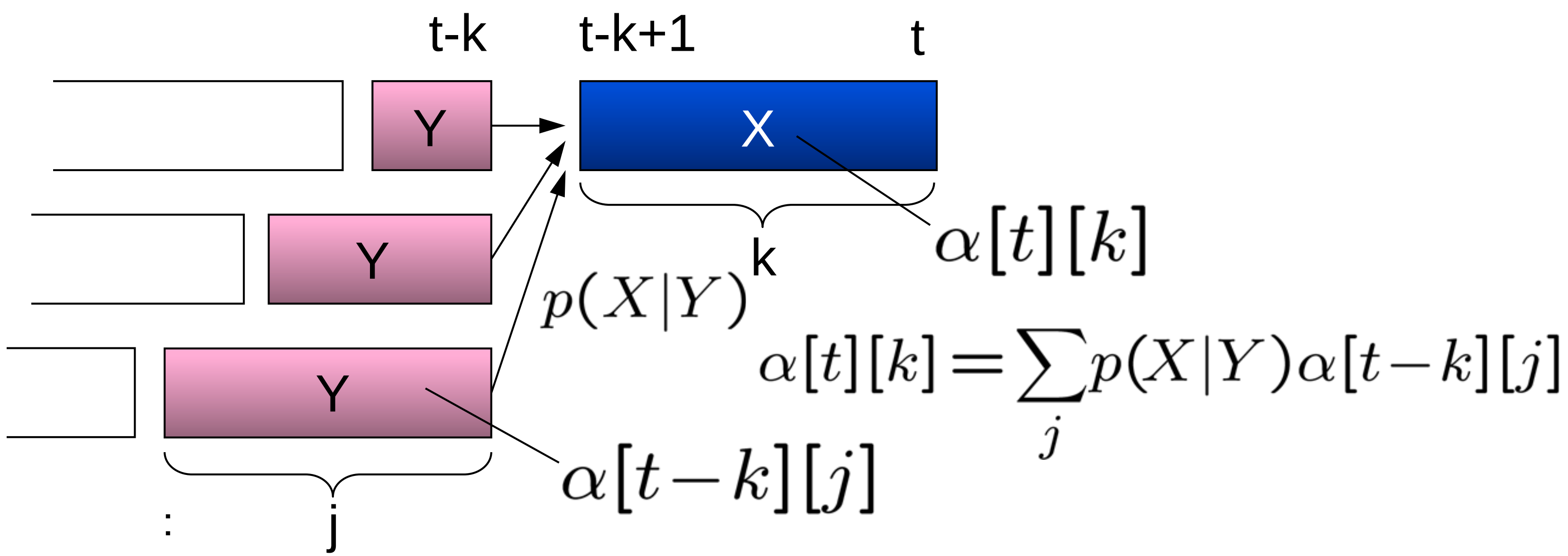
100 神戸 で は 異 人館 街 の 二 十 棟 が 破損 し た 。

200 神戸 で は 異人館 街 の 二 十 棟 が 破損 し た 。

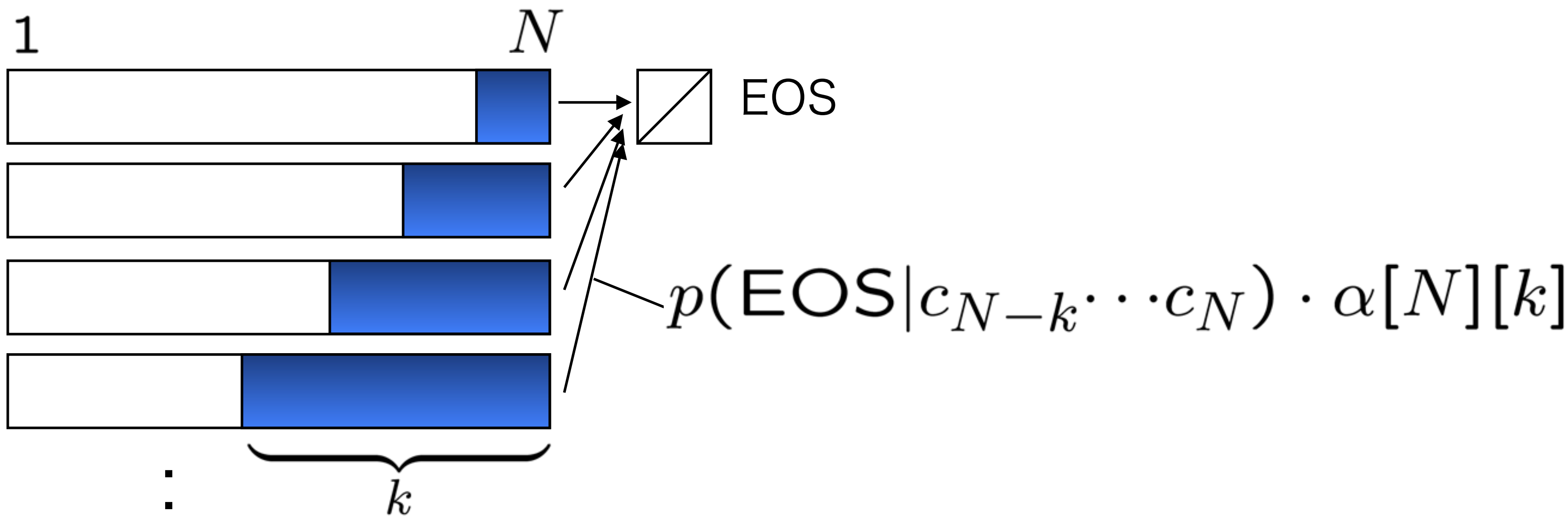
- ギブスサンプリングを繰り返すごとに、単語分割とそれに基づく言語モデルを交互に改善していく。

動的計画法による推論

- words(s)~p(w|s,Θ) : 文sの単語分割のサンプリング
- 確率的Forward-Backward (Viterbiだとすぐ局所解)
  - Forwardテーブル $\alpha[t][k]$ を用いる
  - $\alpha[t][k]$: 文字列 $c_1 c_2 \cdots c_t$ が、時刻tからk文字前までを単語として生成された確率
    - それ以前の分割について周辺化…動的計画法で再帰

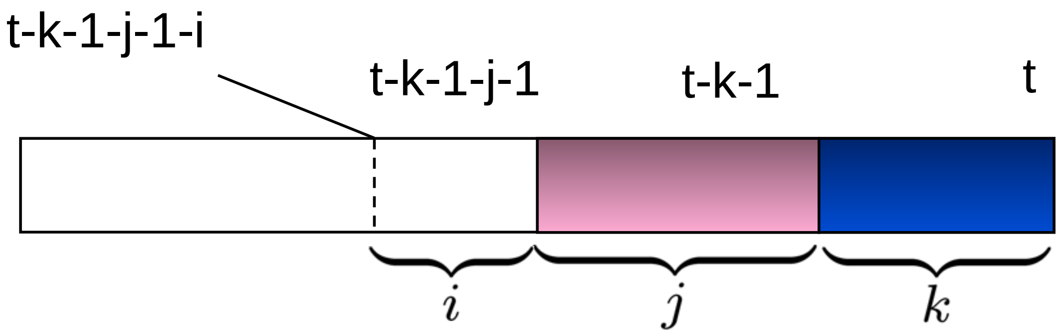


動的計画法によるデコード



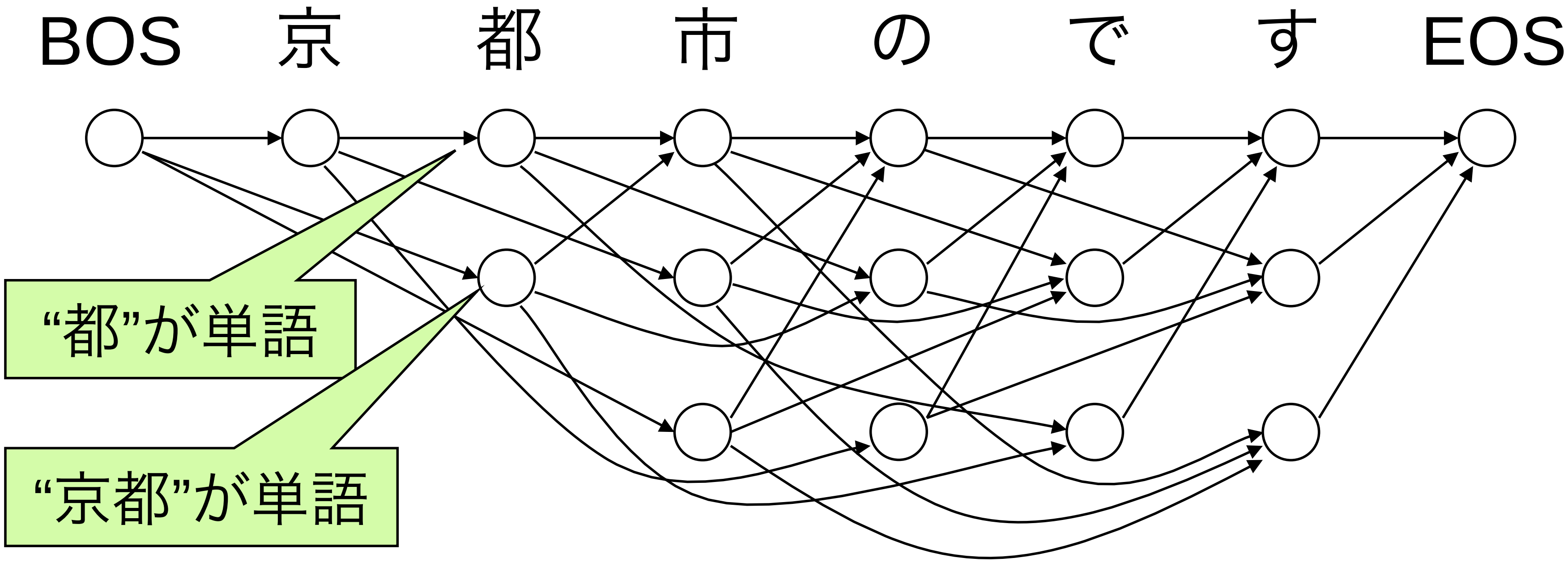
- $\alpha[N][k]$=文字列の最後のk文字が単語となる文字列確率なので、EOSに接続する確率に従って後ろからkをサンプル
- $c_{N-k}\cdots c_N$ が最後の単語だとわかったので、$\alpha[N-k-1][k']$を使ってもう一つ前の単語をサンプル
- 以下文頭まで繰り返す

動的計画法による推論 (トライグラムの場合)



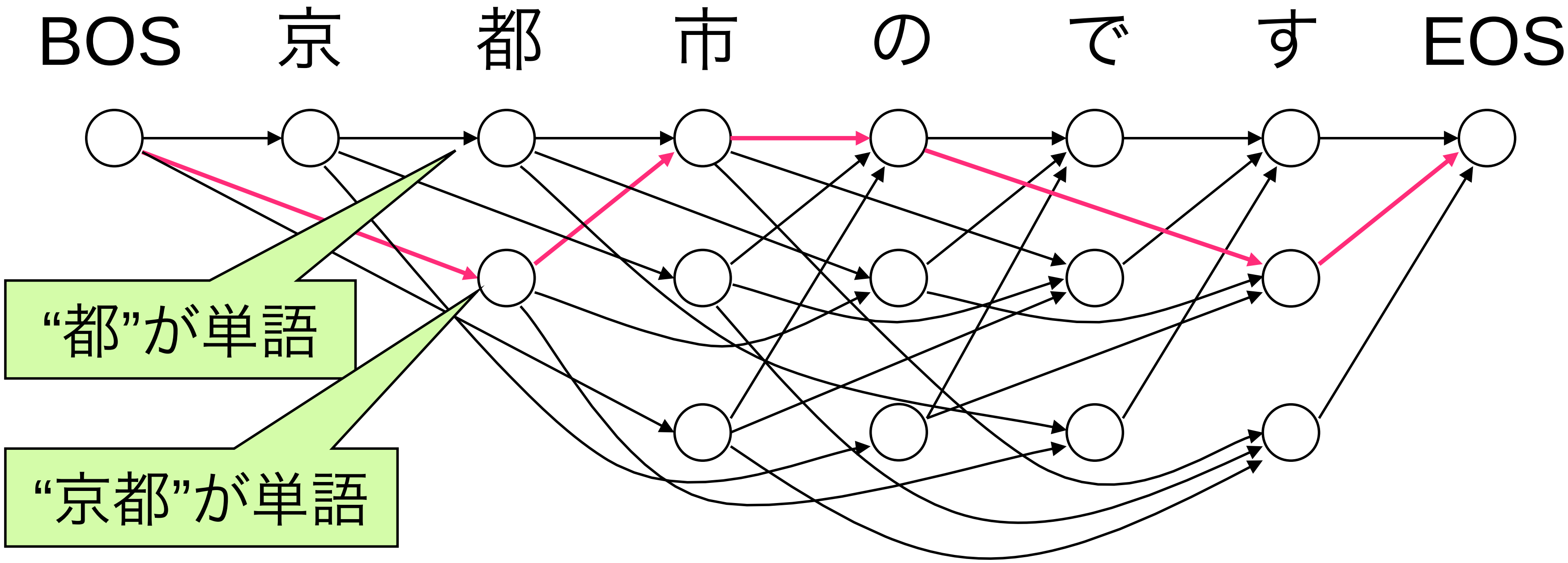
- トライグラムの場合は、Forward 変数として $\alpha[t][k][j]$ を用いる
  - $\alpha[t][k][j]$: 時刻tまでの文字列のk文字前までが単語、さらにそのj文字前までが単語である確率
  - 動的計画法により、$\alpha[t-k-1][j][i]$ $(i = 0 \cdots L)$ を使って再帰
    - プログラミングが超絶ややこしい ;_;
      (文字列は有限なので前が存在しないことがある)

グラフィカルモデル表現 (バイグラム)



- Semi-Markovモデル(Sawaragi&Cohen 04など)と同じ
  - 状態遷移確率が、階層言語モデルで与えられる
  - 正しいパスが何かは未知 (HMMと同じ)
- 動的計画法によって、確率の高いパスをサンプリング

グラフィカルモデル表現 (バイグラム)



- Semi-Markovモデル(Sawaragi&Cohen 04など)と同じ
  - 状態遷移確率が、階層言語モデルで与えられる
  - 正しいパスが何かは未知 (HMMと同じ)
- 動的計画法によって、確率の高いパスをサンプリング

実験: 日本語＆中国語コーパス

- 京大コーパス＆SIGHAN Bakeoff 2005 中国語単語分割公開データセット
- 京大コーパスバージョン4
  - 学習: 37,400文、評価: 1000文(ランダムに選択)
- 日本語話し言葉コーパス: 国立国語研究所
- 中国語
  - 簡体中国語: MSRセット, 繁体中国語: CITYUセット
  - 学習: ランダム50,000文、評価: 同梱テストセット
- 学習データをそれぞれ2倍にした場合も同時に実験

京大コーパスの教師なし形態素解析結果

一方 、 村山富市 首相 の 周囲 に も 韓国 の 状況 や 立場 を 知 る 高官 は い ない 。
日産自動車 は 、 小型 乗用車 「 ブルーバード 」 の 新 モデル ・ Ｓ Ｖ シリーズ ５ 車種 を １ ２ 日 から 発売 した 。
季刊 誌 で 、 今 月 三 十 日 発行 の 第一 号 は 「 車いすテニス 新世代 チャンピオン 誕生 — 斎田悟司 ジャパン カップ 松本 、 平和 カップ 広島 連覇 」 「 フェスピック 北京大会 — 日本 健闘 メダル 獲得 総数 ８ ８ 個 」 「 ジャパン パラリンピック — 日本 の 頂点 を 目指 す 熱い 闘い 」 など の 内容 。
整備新幹線 へ 投入 する 予算 が あ る の なら 、 在来 線 を 改良 する などして、 高速 化 を 推進 し 輸送力増強 を 図 れ ば よい 。
国連 による 対 イラク 制裁解除 に 向け 、 関係 の深い 仏 に 一層 の 協力 を 求め る の が 狙い とみられる 。
この 日 、 検査 され た の は ワシントン州 から 輸出 され た 「 レッド デリシャス 」 、 五 ニ トン 。

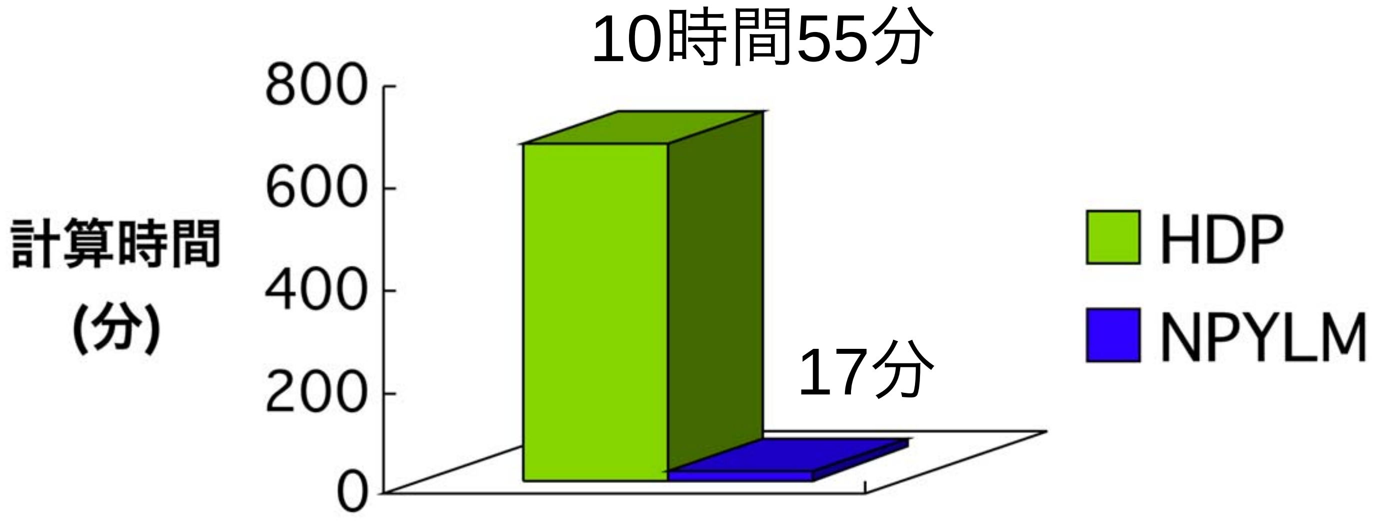
ビタビアルゴリズムで効率的に計算可能
(先行研究では不可能)

“正解”との一致率 (F値)

モデル	MSR	CITYU	京大
NPY(2)	0.802 (51.9)	**0.824** (**126.5**)	0.621 (23.1)
NPY(3)	**0.807** (**48.8**)	0.817 (128.3)	**0.666** (**20.6**)
NPY(+)	0.804 (**38.8**)	0.823 (**126.0**)	**0.682** (**19.1**)
ZK08	0.667 (—)	0.692 (—)	—

- NPY(2),NPY(3)＝NPYLM 単語バイグラムorトライグラム+文字∞グラム
  - NPY(＋)はNPY(3)でデータを2倍にしたもの
- 中国語: ZK08＝(Zhao&Kit 2008)での最高値と比べ、大きく改善
  - ZK08はヒューリスティックな手法をさらに混合したもの

計算時間の比較



- HDP(Goldwater+ ACL 2006): 学習データのすべての文字について1文字ずつサンプリング
  – モデルは単語2グラムのみ (文字モデルなし)
- NPYLM: 文毎に動的計画法により効率的にサンプリング
  – 単語3グラム-文字∞グラムの階層ベイズモデル

日本語話し言葉コーパス (国立国語研究所)

うーんうんなってしまうところでしょうねへーあーでもいいいいことですよねうーん
うーん自分にも凄くプラスになりますものねそうですねふーん羨ましいです何かうーん精神的にもう子供達に何かこう支えられるようなうーものってやっぱりあるんですよやってるとうーんうーんうーん
うーん長くやってればそんなものがうんうんそうでしょうねたくさんやっぱりありますねうんうーんなるほど…

NPYLM

うーん うん なって しまう ところ でしょう ね へー あー でも いい いい こと ですよねうーん
うーん 自分 に も 凄く プラス に なり ます もの ね そう です ね ふーん 羨ましい です 何か うーん 精神的にもう 子供達に何か こう 支えられる ような うー もの って やっぱり ある んです よ や って る と うーん うーん うーん うーん 長く や って れば そんな もの が うん うん そう でしょう ね たくさん やっぱり あり ます ね うん うーんなるほど…

「源氏物語」の教師なし形態素解析

しばしは夢かとのみたどられしを、やうやう思ひしづまるにしも、さむべき方なくたへがたきは、いかにすべきわざにかとも、問ひあはすべき人だになきを、忍びては参りたまひなんや。若宮の、いとおぼつかなく、露けき中に過ぐしたまふも、心苦しう思さるるを、とく参りたまへ』など、はかばかしうも、のたまはせやらず、むせかへらせたまひつつ、かつは人も心弱く見たてまつるらむと、思しつつまぬにしもあらぬ御気色の・・・・

NPYLM

しばし は 夢 か と のみ たど られ し を 、 やうやう 思ひ しづま る に
しも 、 さむ べき 方 な く たへ がた き は 、 いかに す べき わざ に か
と も 、 問ひ あは す べき 人 だに な き を 、 忍びて は 参り たまひ な
ん や 。若 宮 の 、 いと おぼつかな く 、 露け き 中に 過ぐし たまふ も
、 心 苦し う 思さる る を 、 とく 参り たまへ 』 など 、 はかばかしう
も 、 のたまはせ やら ず 、 むせ かへ ら せ たまひ つつ 、 かつ は 人も
心 弱 く 見 たてまつ る ら む と 、 思しつつ ま ぬ に しも あら ぬ 御
気色 の・・・・

アラビア語教師なし形態素解析

- Arabic Gigawords から40,000文 (Arabic AFP news)

الفلسطينيبسببتظاهرةلانصارحركةالمقاومةالاسلاميةحماس.

Google translate:
"Filstinebsbptazahrplansarhrkpalmquaompalaslamiphamas."

NPYLM

الفلسطيني بسبب تظاهرة ل انصار حركة المقاومة الاسلامية حماس .

Google translate:
"Palestinian supporters of the event because of the Islamic Resistance Movement, Hamas."

“Alice in Wonderland”の解析

first,shedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspedup
onherknee,andthebrighteagereyeswerelookingupintohersshecouldhearthevery
tonesofhervoice,andseethatqueerlittletossofherheadtokeepbackthewanderingh
airthatwouldalwaysgetintohereyesandstillasshelistened,orseemedtolisten,thew
holeplacearoundherbecamealivethestrangecreaturesofherlittlesister'sdream.the
longgrassrustledatherfeetasthewhiterabbithurriedbythefrightenedmousesplashe
dhiswaythroughtheneighbouringpoolshecouldheartherattleoftheteacupsasthema
rchhareandhisfriendssharedtheirneverendingmeal,andtheshrillvoiceofthequeen…

first, she dream ed of little alice herself ,and once again the tiny hand s were
clasped upon her knee ,and the bright eager eyes were looking up into hers --
shecould hearthe very tone s of her voice , and see that queer little toss of
herhead to keep back the wandering hair that would always get into hereyes --
and still as she listened , or seemed to listen , thewhole place a round her
became alive the strange creatures of her little sister 'sdream. thelong grass
rustled ather feet as thewhitera bbit hurried by -- the frightened mouse splashed
his way through the neighbour ing pool -- shecould hearthe rattle ofthe tea cups
as the marchhare and his friends shared their never -endingme a l ,and the …

まとめ

- ベイズ単語nグラム-文字nグラムを階層的に統合した言語モデルによる、教師なし形態素解析
  - 動的計画法＋MCMCによる効率的な学習
- あらゆる自然言語に適用できる
  - データに自動的に適応、「未知語」問題がない
  - 識別学習と違い、学習データをいくらでも増やせる
  - 話し言葉、ブログ、未知の言語、古文、…
- あらゆる言語の文字列から直接、「単語」を推定しながら言葉のモデルを学習する方法ともみなせる

実装

- 数万〜数十万文 (数百万〜数千万文字)の学習テキストに対してGibbsサンプリングを繰り返すため、高速な実装が不可欠
  - MATLABやRでは計算が追いつかない
- C++&Cで実装, 6000行程度
  - 解析速度は100〜200文/秒 (10ms/文以下)
  - 1つの文を解析するのに、nグラム確率を40000回程度計算する必要
  - 階層的データ構造の動的なアップデート
  - 学習時間: 10〜20時間程度

展望

- 教師あり学習と異なり、学習データをいくらでも増やせる ➡ 学習の高速化、並列化
  - HDP-LDAのGibbsの並列化 (Welling+, NIPS 2007-2008) が適用可能
- 識別学習との融合による半教師あり学習
  - Loglinearの枠組で統合するにも、生成モデルが必要
    - これまで、生成モデルが存在しなかった
    - 提案法は、CRFのForward-Backwardの教師なし版のようなもの
    - POS Tagging: CRF+HMM (鈴木,藤野+ 2007)で提案

最近の関連研究

- 音声認識での音素モデルの分割
  - Neubig+, 情報処理学会 *SLP82*, 2010 July
- 統計的機械翻訳への適用
  - Nguyen+ (CMU LTI), *COLING 2010*
- [参考1] 文法構造を仮定した分割
  - Johnson&Goldwater, *NAACL 2009*
- [参考2] Log-Linearモデルとしての定式化
  - Poon+, *NAACL 2009*

音声認識への適用 (Neubig+ 2010)

- 音声認識の目標：音声信号(音素列) $\mathbf{a}$ から単語列 $\mathbf{w}$ を復元すること

$$p(\mathbf{w}|\mathbf{a}) \propto p(\mathbf{a}|\mathbf{w})p(\mathbf{w})$$ (ベイズの定理)

  - 問題： $p(\mathbf{w})$ で会話文の「単語」を膨大な人手で作成
  - 人手の単語分割なしに、「単語」を自動認識できないか? [未知の言語の会話文の解析にも有効]

学習された例: (WFST/OpenFST上で動作)

「mo-shiagetemasukedmo」→
「mo-shiage te ma sukedomo」(申し上げ て ま すけども)
「kyo-moshikkariyarimasuyoto」→
「kyo- mo shikkari yarimasu yoto」
(今日 も しっかり やります よと)

学習された単語

助詞

	例	頻度（順位）
no	国民のための	1052 (1)
ni	地域に密着	830 (2)
to	韓国との対話	685 (5)

話し言葉表現

音素	例	頻度（順位）
yu:	と言うのは	324 (19)
e:	えー、それでは	202 (28)
desune	それはですね	94 (65)

rimasukeredomo, mo:shiage, yu:fu:ni

サブワード

	例	頻度（順位）
ka	強化、カイカク	713 (3)
to:	答弁、とー言う	204 (27)
sai	制裁、国際	94 (65)

内容語・機能語

音素	例	頻度（順位）
koto	こと	189 (32)
omo	思って	56 (109)
hanashi	話	23 (242)

jo:kyo:, kangae, chi:ki, toki, shiteki

統計的機械翻訳への適用

- 統計的機械翻訳＝「英語←→中国語」のような文対から、完全に自動的に翻訳規則を学習

我们想要张靠窗户的桌子 。 We want to have a table near the window. …

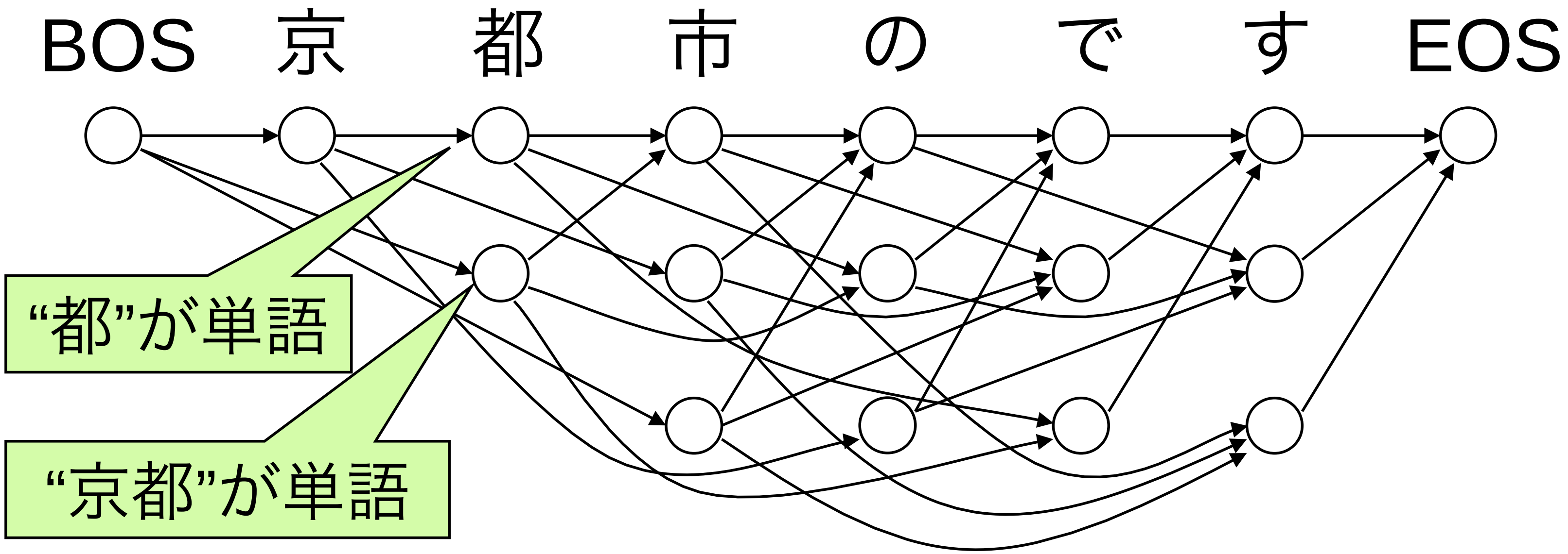
  - 原言語の単語分割を、翻訳に最適化するべき
- 先行研究: “Bayesian Semi-Supervised Chinese Word Segmentation for Statistical Machine Translation”, Xu+, *COLING 2008*.
- 本研究の翻訳への拡張: “Nonparametric Word Segmentation for Machine Translation”, Nguyen, Vogel, & Noah Smith, *COLING 2010*

[参考1] 文法構造による分割

(Johnson&Gold-
water, NAACL 2009)

- 文の文字列が、次のような規則(一部)で再帰的に生成されたと仮定:
  - Sentence→Colloc+
  - Colloc→Word+
  - Word→SyllableIF
  - Word→SyllableI (Syllable) (Syllable) SyllableF
  - Syllable→Onset Rhyme
  - SyllableIF→OnsetI RhymeF …
- 子供の発話コーパス (Bernstein-Ratnerコーパス)に対して正解率88%程度で最高精度
  lUkD*z6b7wIThIzh&t→lUk D*z 6 b7 wIT hIz h&t
  (look there's a boy with his hat)
  - ただし、平均9.79文字/文の短い文/音素列のみ

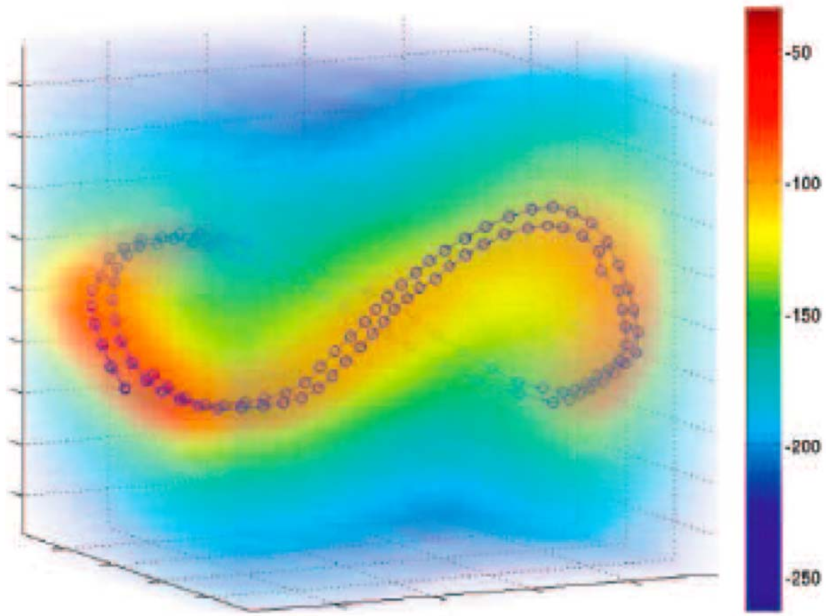
[参考2] 対数線形モデルによる分割



- グラフィカルモデルのパスの重みは、必ずしも確率になっていなくても良い→Log-linearモデル
  - Prefix, Suffix等の重み付け
- (Poon+ 2009): 単語分割は既知で、その内部を分割
  - アラビア語: “wvlAvwn”→“w-vlAv-wn” (最大14文字)

学習はcontrastive estimation

連続系のモデリングの場合

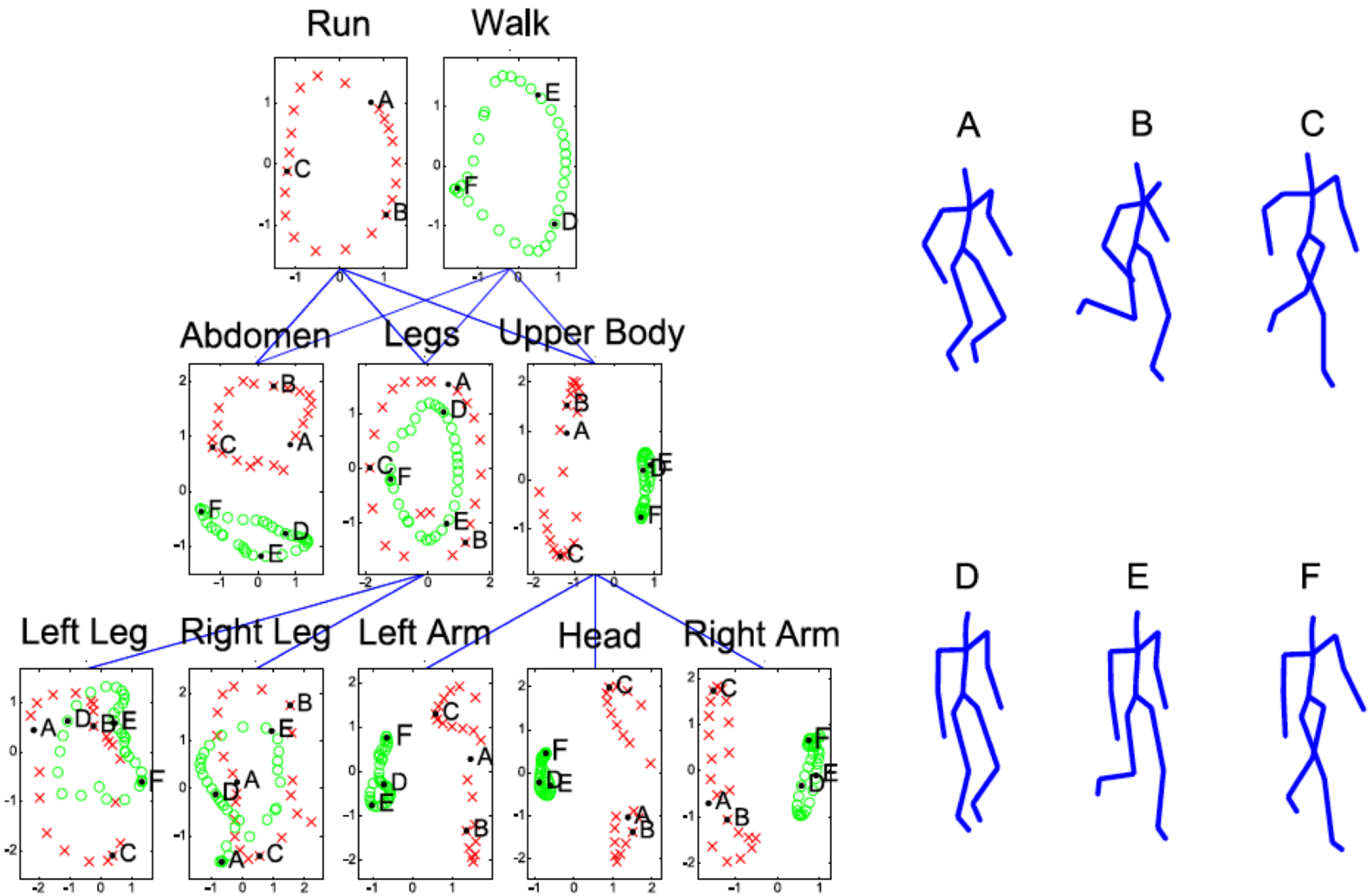


- Wang+, “Gaussian process dynamical model”, PAMI 2008/NIPS 2005

$$p(\mathbf{X}, \mathbf{Y}, \bar{\alpha}, \bar{\beta}, \mathbf{W}) = \underbrace{p(\mathbf{Y} \mid \mathbf{X}, \bar{\beta}, \mathbf{W})}_{\text{潜在座標→観測値モデル}} \underbrace{p(\mathbf{X} \mid \bar{\alpha})}_{\text{潜在座標の非線形ARモデル}} p(\bar{\alpha})\, p(\bar{\beta})\, p(\mathbf{W}).$$

連続系のモデリング (2)

- Lawrence&Moore, “Hierarchical Gaussian process Latent variable models”, ICML 2007.



文献

本研究：

- “Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor language modeling”, ACL 2009.
- 「ベイズ階層言語モデルによる教師なし形態素解析」情報処理学会 NL-190, 2009.

関連研究：

- 「ベイズ推論を用いた連続音声からの言語モデル学習」Graham Neubig+, 情報処理学会 SLP-82, July 2010.
- “Nonparametric Word Segmentation for Machine Translation”, ThuyLinh Nguyen+, *COLING 2010.*
- “Improving nonparametric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars”, Mark Johnson & Sharon Goldwater, NAACL 2009.
- “Unsupervised Morphological Segmentation with Log-Linear Models”, Hoifung Poon, Colin Cherry, Kristina Toutanova, NAACL 2009.

まとめ

- 言語モデル・・・・自然言語の教師なし学習
  - nグラムとは限らない
  - 形態素解析、構文解析など、NLPの殆どのタスクを内包
- HPYLM, VPYLM・・・・nグラム(∞グラム)のベイズモデル
  - 従来と違い、他の確率モデルの部品にできる
  - 応用: 教師なし形態素解析、半教師あり形態素解析
- 自然言語の教師なし学習は、未だ急速に発展中
  - 半教師あり学習のためにも、さらに研究が必要
  - 個人的には、より深い統計に基づくノンパラメトリックベイズ法に興味

終わり

- ご清聴ありがとうございました。