ノンパラメトリックベイズ法による言語モデル

持橋大地
統計数理研究所 モデリング研究系
daichi@ism.ac.jp

2012-3-15 (木), 統数研

Overview

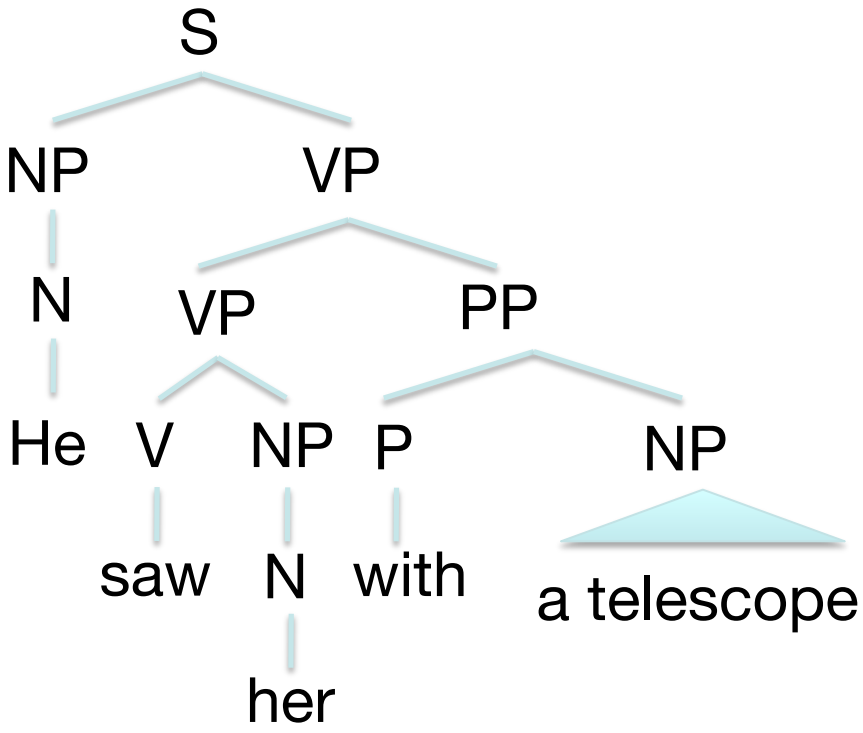
- 統計的自然言語処理・言語モデルとは
- ノンパラメトリック・ベイズ法とは (イントロ)
  - Dirichlet分布
  - Dirichlet過程
  - 階層Dirichlet過程
  - Chinese Restaurant Process (CRP)
- 階層Pitman-Yor過程に基づくnグラム言語モデル
- 階層言語モデルによる教師なし形態素解析

言語学と統計的自然言語処理

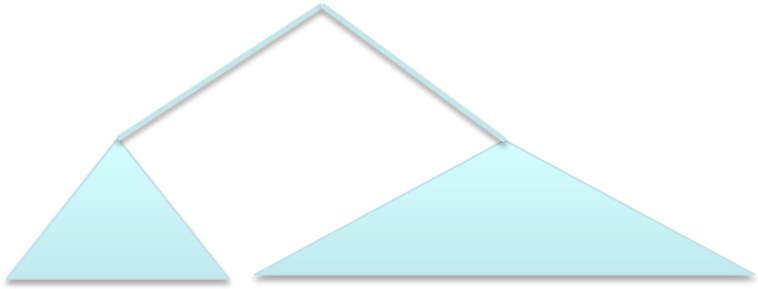
- 言語の研究→言語学科に行けばよいか?
- いわゆる「言語学」＝手で書いたルールの固まり！
- 例：構文解析

S→NP VP
VP→V PP
VP→V NP
PP→P NP
NP→DET N
NP→N

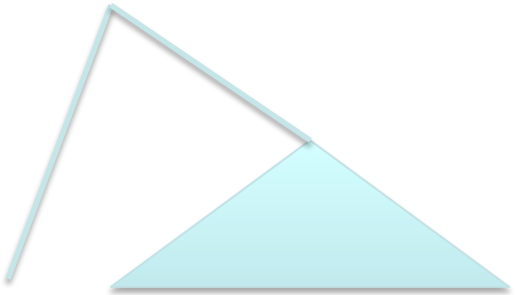
文法ルール



言語学と統計的自然言語処理 (2)



He saw her with a telescope

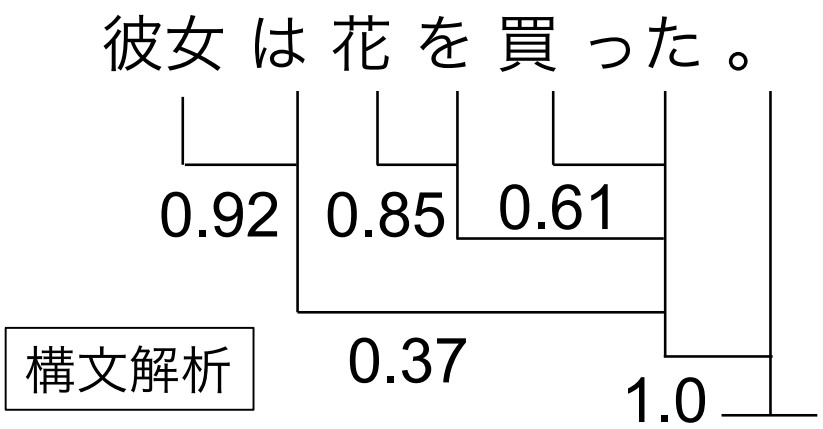


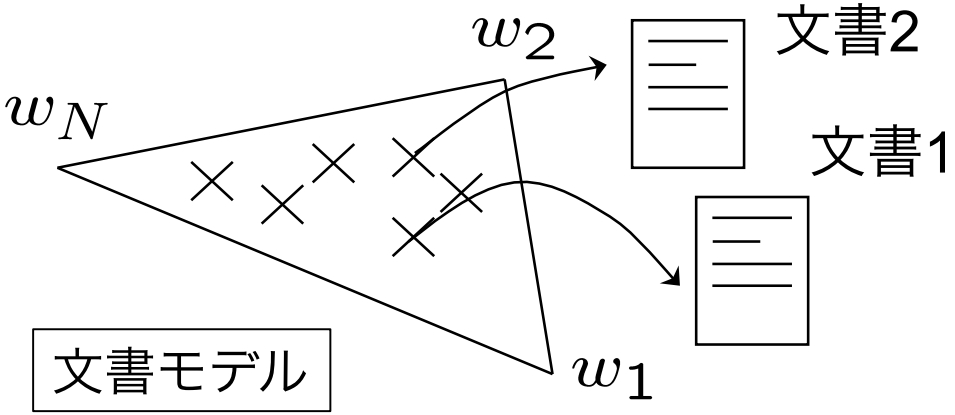
He saw her with a hat

- 解釈が名詞によってなぜ違うのか?
- 古典的な言語学では答えが出せない・
  そもそも主観的
  → 確率モデル・統計学として数学的に
  考え直す必要！
  - cf. 中世の天動説から地動説の数学理論へ

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
  - Webの出現、大量の電子テキスト
- 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、・・・・





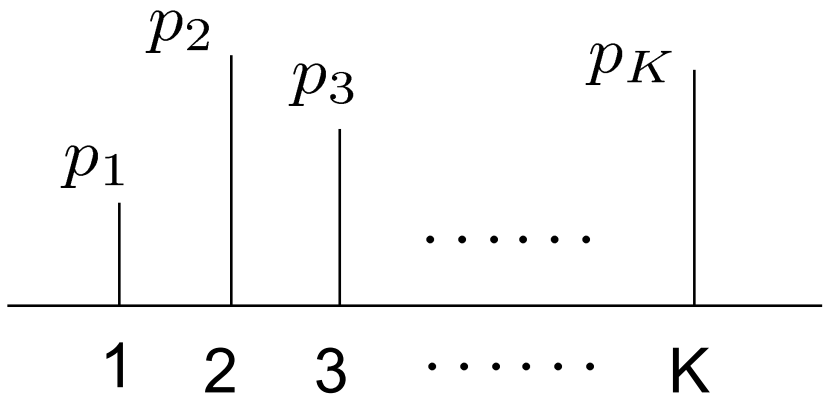
統計的言語モデル

- 統計的自然言語処理の最も基本的なモデル
- 単語列 $\mathbf{w} = w_1 w_2 \cdots w_k$ に対し、その確率 $p(\mathbf{w})$ を最大にする確率モデルを学習
  - 木構造やMarkovモデルなど
  - 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…
  - 自然言語処理のほとんどすべての問題を含む

多項分布 (離散分布)

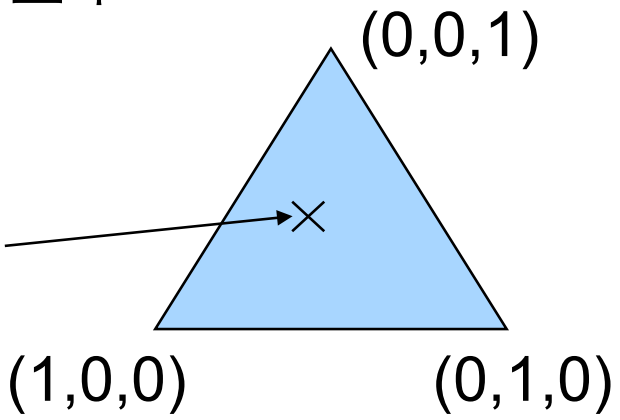


$$\sum_{k=1}^{K} p_k = 1,$$

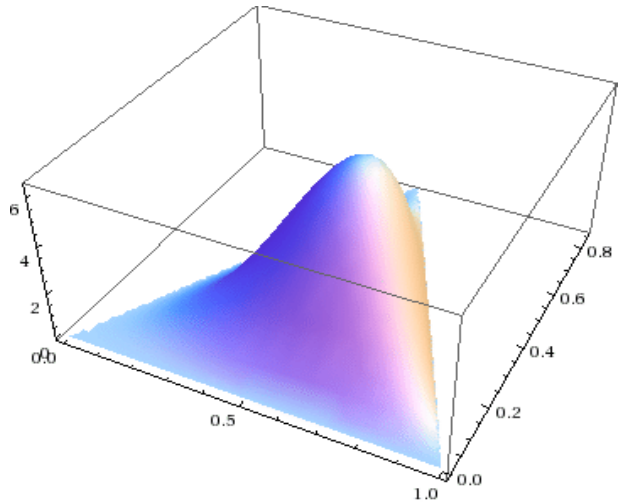
$$\forall k,\ p_k \geq 0$$

- 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



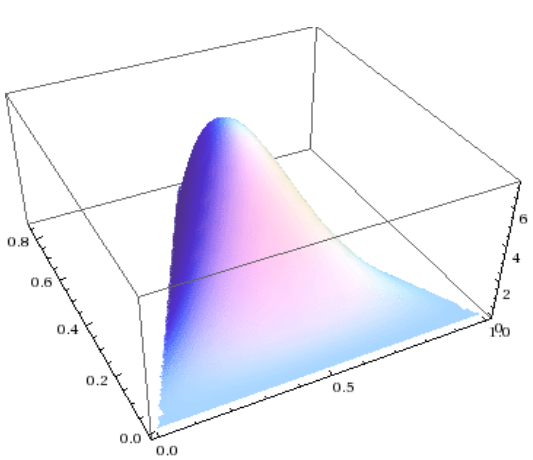
$$\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} p_k^{\alpha_k - 1}$$

パラメータ:

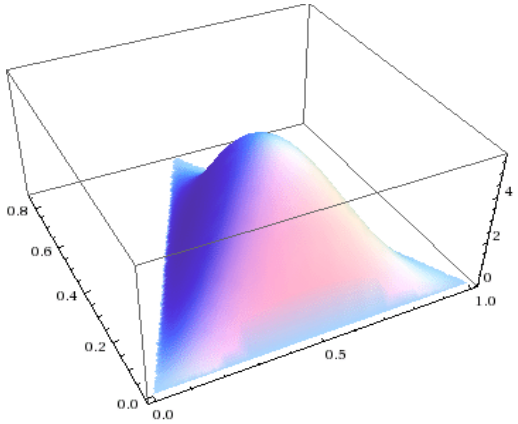
$$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$$

- 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」　： $\alpha = \sum_k \alpha_k$

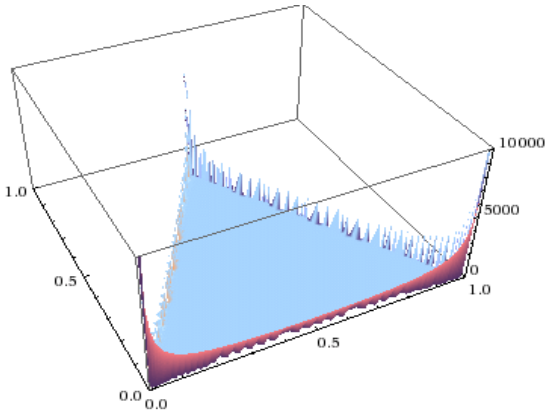
ディリクレ分布 (2)



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



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



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

- $\alpha_k > 1$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　　$\alpha \ll 1$
    ($\alpha = 0.1 \sim 0.0001$ 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: $p(\mathbf{p}|X) \propto p(X|\mathbf{p})p(\mathbf{p})$

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

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

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

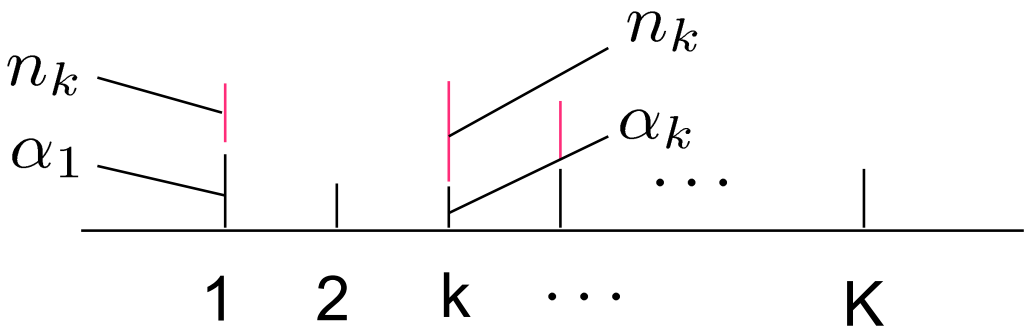
- $\mathbf{p}$の期待値は、

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

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

(出現しなかったアイテムにも正の確率)

ノンパラメトリック・ベイズ法とは

- モデルの複雑さを、データの複雑さに応じて無限に伸縮することのできるベイズ統計モデル
  - 「パラメータがない」という意味ではない
- 簡単な場合の例
  - GMMの混合数
  - HMMの隠れ状態数
  - 文書に存在する意味トピック数
  - 言語の文法的ルールの複雑さ
- 有名なもの: Dirichlet過程 (無限次元Dirichlet分布)

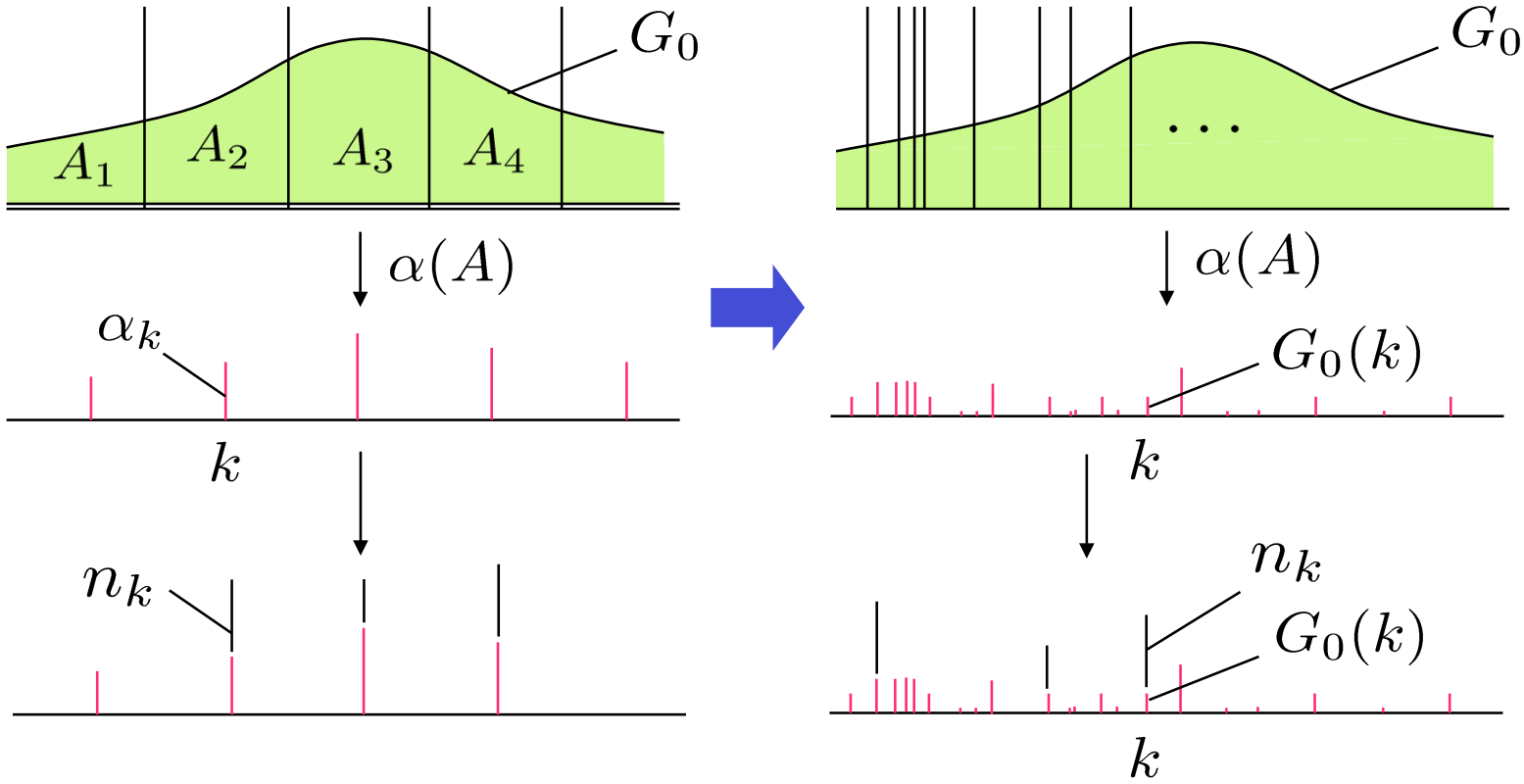
ディリクレ過程

- Dirichlet processとは要するに何?
  → 無限次元ディリクレ分布.
- DPの定義 (Ferguson 1973):

> A stochastic process *P* is said to be a Dirichlet process on $(\mathscr{X}, \mathscr{A})$with parameter α if for any measurable partition $(A_1, \cdots, A_k)$of $\mathscr{X}$ the random vector $(P(A_1), \cdots, P(A_k))$ has a Dirichlet distribution with parameter $(\alpha(A_1), \cdots, \alpha(A_k))$

- どういうこと??

ディリクレ過程 (2)



予測確率:

$$\frac{\alpha_k + n_k}{\alpha + n}$$

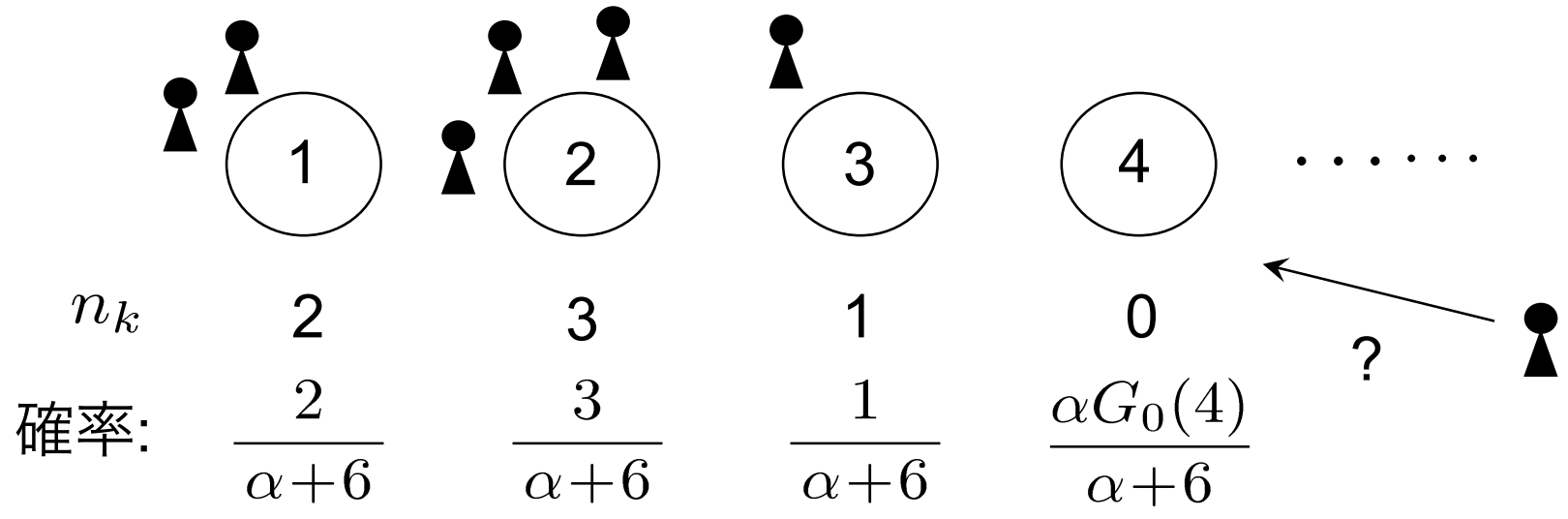
$$\frac{\alpha G_0(k) + n_k}{n + \alpha}$$

Chinese Restaurant Process (CRP)

- 予測確率

$$p(k|X) = \frac{\alpha_k + n_k}{\alpha + n} \text{ (Dirichlet)}, \quad \frac{\alpha G_0(k) + n_k}{\alpha + n} \text{ (DP)}$$

  - ディリクレ分布/過程に従うと、頻度 n_k 高いものはさらに現れやすくなる (rich-gets-richer) → CRP



ディリクレ過程と言語モデル

- ディリクレ過程は、語彙が無限の場合の単語の確率分布ともみることができる

$$p(w|X) = \frac{c(w)}{\alpha+n} + \frac{\alpha}{\alpha+n} G_0(w)$$

  - カウントc(w)が0のどんな未知の単語wでも、$G_0(w) \cdot \alpha/(\alpha+n)$の確率を持つ
- この確率分布は、p(w)‥‥ユニグラムモデル
  - 単語が独立に出現すると仮定している
  - 一般には、前の単語などに強く依存
    - “is going”→to, “united states of”→america など

nグラムモデルのベイズ学習

- 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

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

現在のGoogle カウント

  - 頻度そのままでなく、何か値を足したりする必要!

ディリクレスムージング (MacKay 1994)

- nグラム確率分布 $p(w|h)$ は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

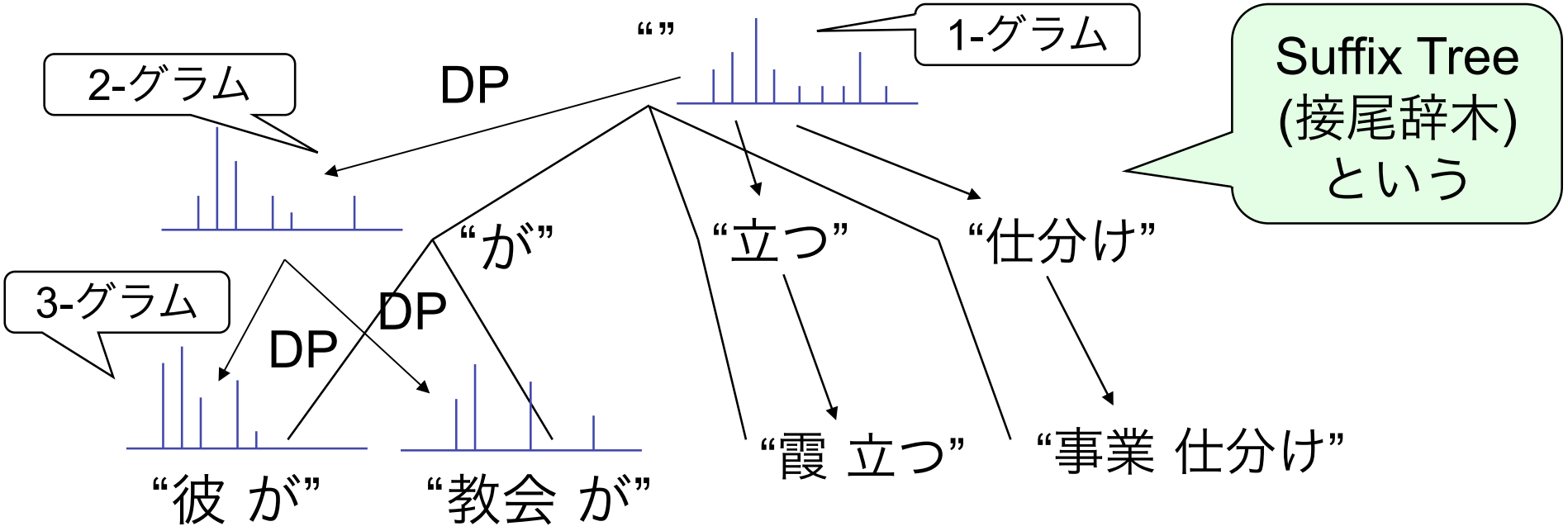
$$p(w|h) = \frac{n(w|h) - D(n(w|h))}{n(h)} + \gamma(h)p(w|h')$$

  - 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?

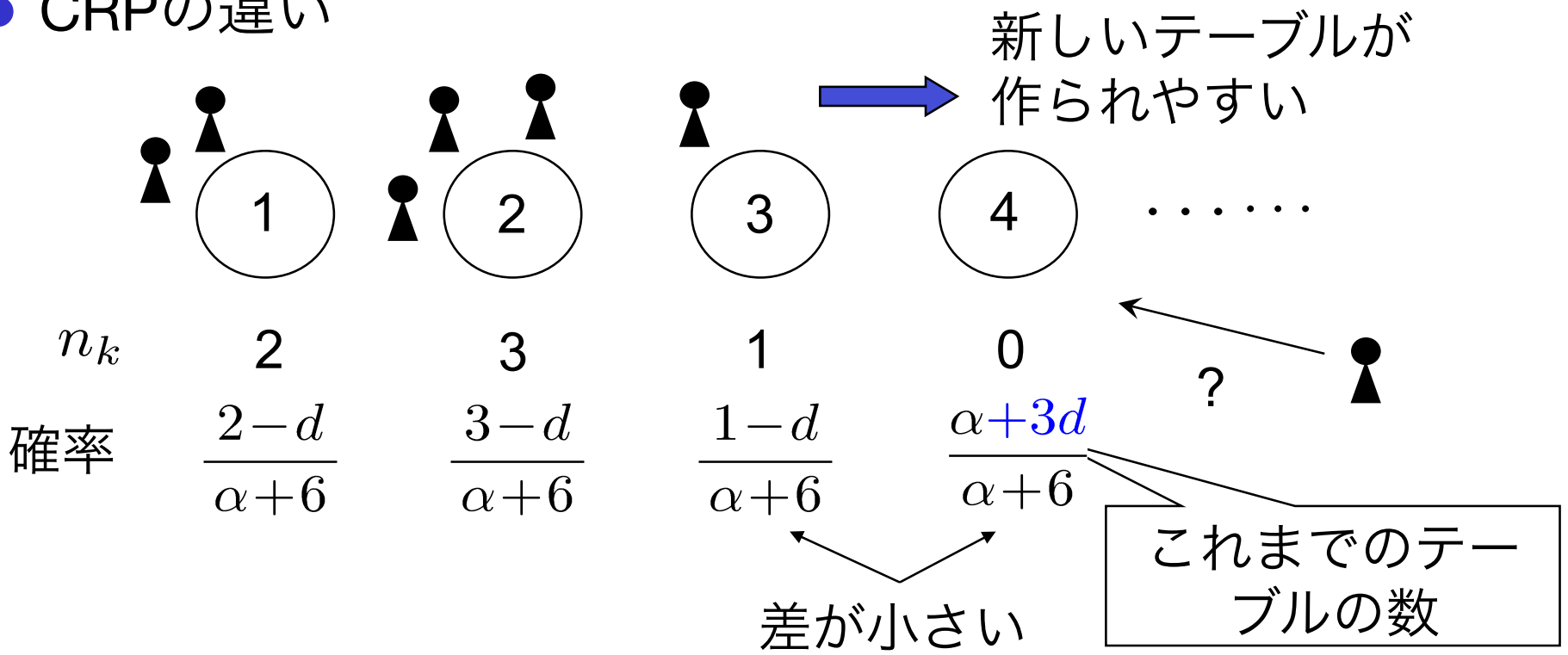
階層ディリクレ過程 (HDP)



- 統計的自然言語処理の広い範囲で、nグラムモデル(=言葉のMarkovモデル)が重要
  - nグラム… 前の(n-1)語に依存して次の語が出現
- nグラム分布を基底測度として、DPで(n+1)グラム分布を生成する

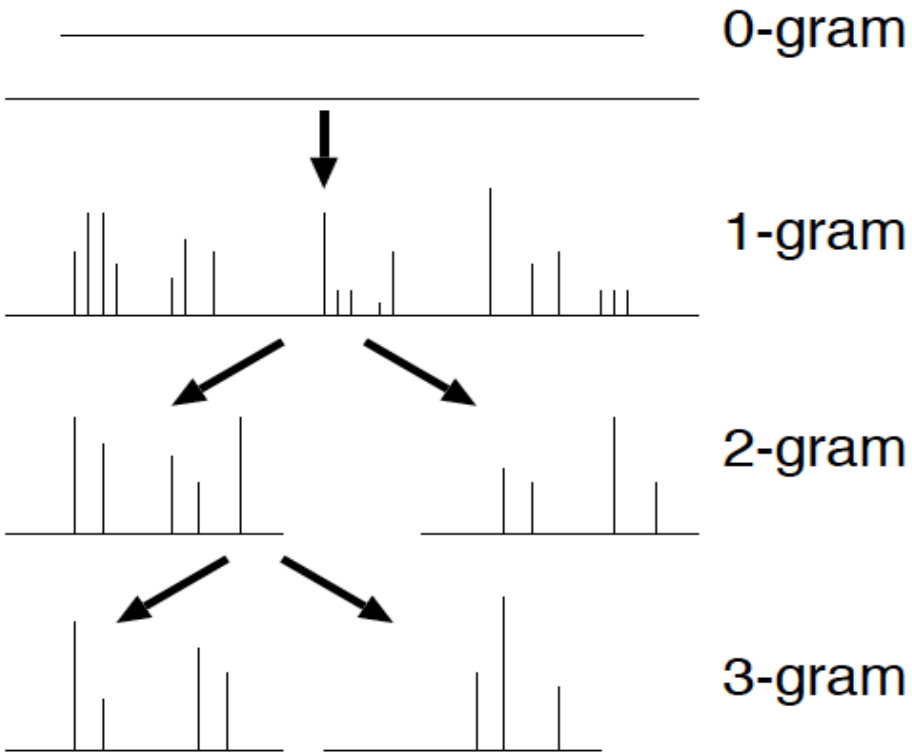
Pitman-Yor過程

- Pitman-Yor過程 (Pitman and Yor 1997): PY(α,d)
  - ディリクレ過程の拡張, Poisson-Dirichlet
  - 新たにディスカウント係数dを持つ
- CRPの違い



階層Pitman-Yor過程 (1)

- 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

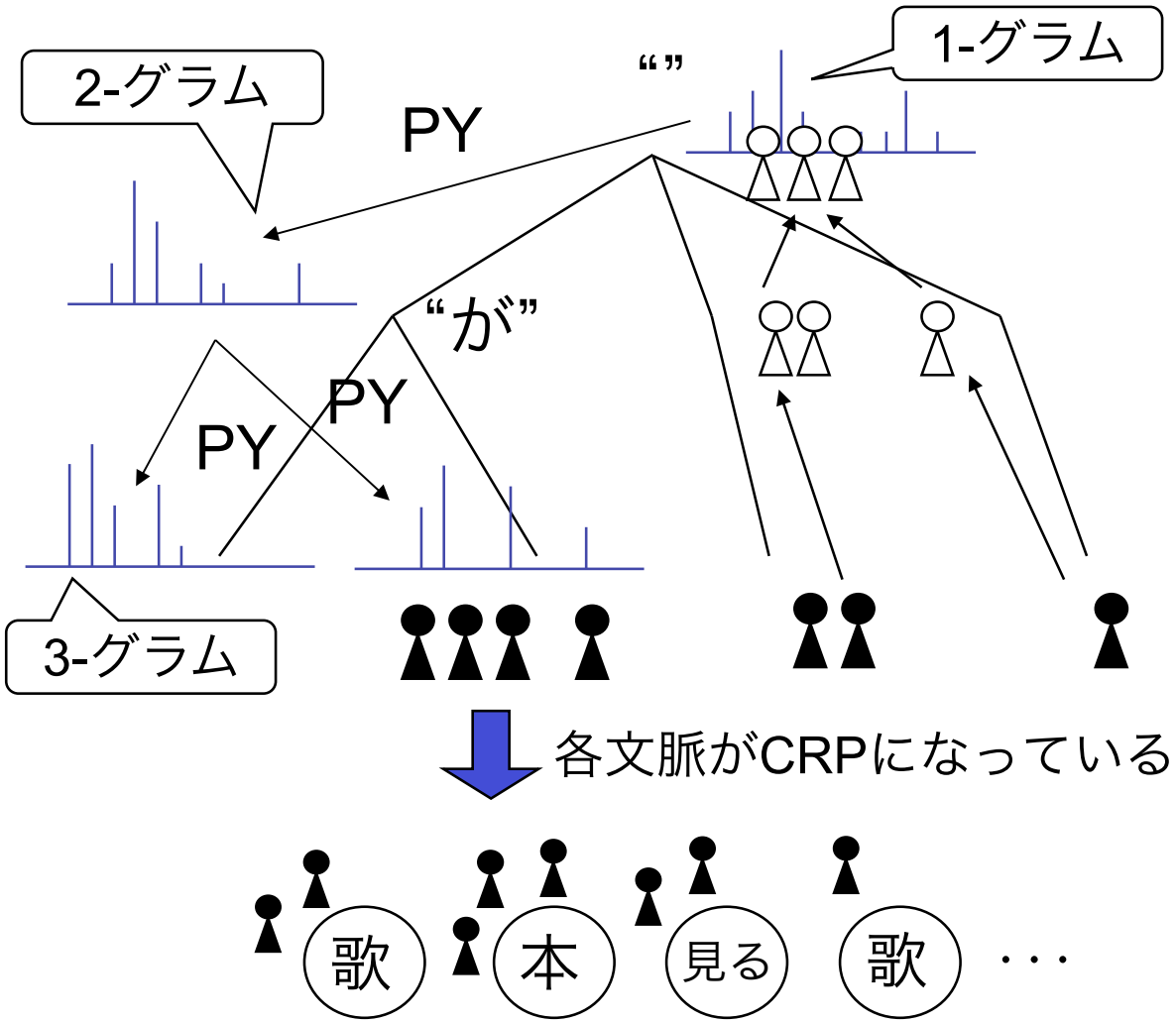


Prior: $(n-1)$ gram

$\downarrow \mathrm{PD}(d_n, \theta_n)$

Posterior: (n) gram

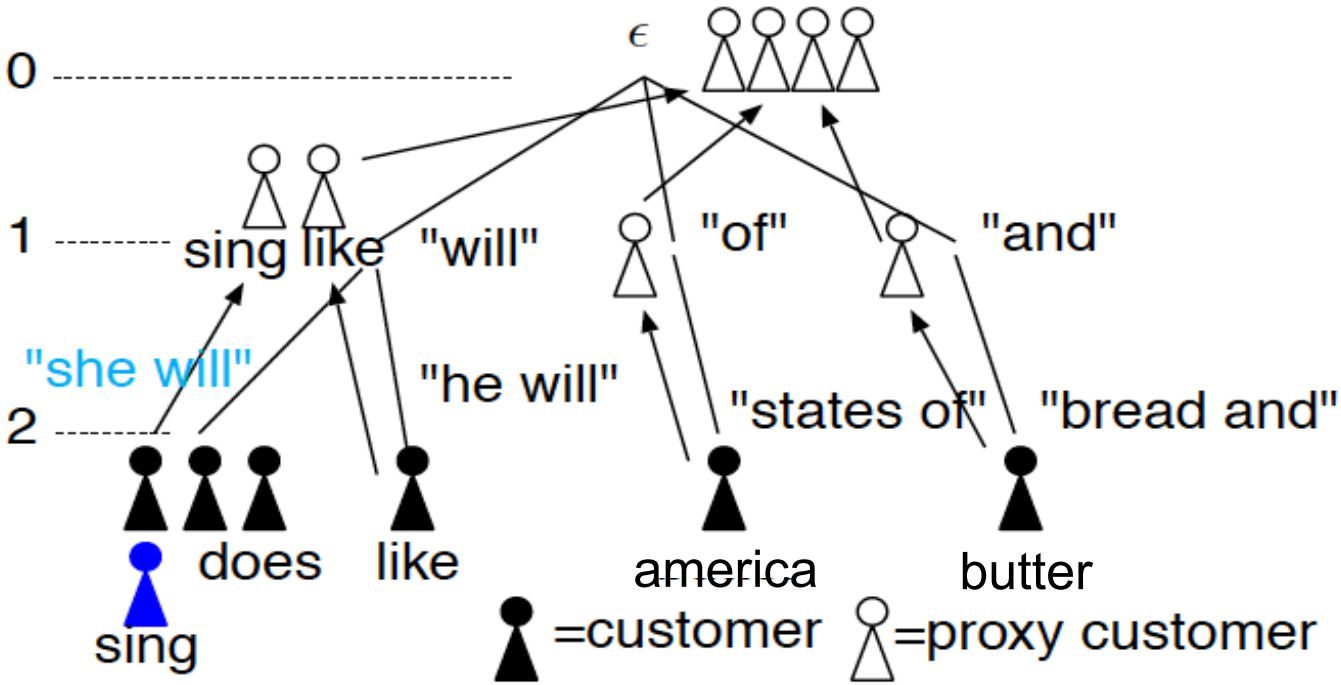
階層Pitman-Yor過程 (2)



- 各nグラム文脈hでの単語出現がCRPに従う
- カウントを複数のテーブルで表現
  →テーブルが増えたとき、客のコピー(代理客)を(n-1)グラム文脈に送る
  - sing a|song→ a|song→ song

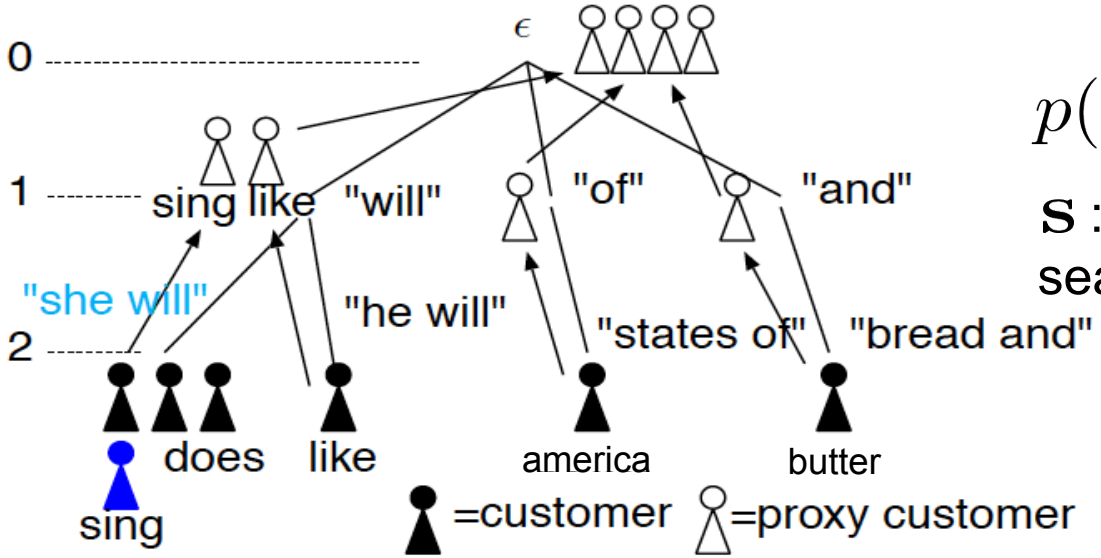
階層CRP表現

- 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の予測確率 (再掲)

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

- 文脈hの下での単語wの予測確率
  - 一つ短い文脈h'での同様な予測確率との混合
- $t_{hw} \equiv 1$のとき、Kneser-Ney スムージングと一致
  - 実際には t_{hw}はO(log(n))で増えることが示されている

階層言語モデルによる教師なし形態素解析

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, 同志社内はどうですか。“
 | mecab -O wakati
やあ こんにちは , 同志社 内 は どう です か 。
(やあこん にち は , 同志 社内 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|、|…

形態素解析 (3)

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、‥

קפיטליזסהיאשיטהכלכליתוחברתיתשהתפתחהבאירופהבין המאהההשעשרהוהמאההתשעעשרה,המבוססתבעיקרהעלהזכותשלפרטיםוקבוצותלבעלותפרטיתעלה רכושולשימושבובאופןחופשי,תוךהסתמכותעלאכיפתזכויותהקניין באמצעותהרשות.

- 古語や、未知の言語の文に関しては
  そもそも何が単語なのかわからない！
- 世界の他の言語でも同様の問題

中国語:


Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.tsinghua.edu.cn/qhdwzy/board_xxgk2.jsp?board=12&bid2=1201&pagen
Google
清華大学
Tsinghua University
学校概况 | 院系设置 | 管理机构 | 科学研究 | 教师队伍 | 人才培养 | 招生就业 | 虚拟校园
学校沿革
历任校长
统计资料
校长致辞
喜迎百年华诞 再铸新的辉煌
——2011年新年献辞
●校长 顾秉林
一世纪沧桑砥砺，一百年春华秋实。此时此刻，2010年的余晖散去，2011年的曙光在前！此时此刻，清华正在送走她的第一个百年，迈向新百年的征程！在辞旧迎新、钟声回荡之际，请允许我代表学校，向清华全体同学、教职员工、离退休人员和海内外广大校友，向长期关心支持清华发展的各界人士，致以崇高的敬意和新年的祝福！
一百年来，清华大学的发展始终与国家民族的命运休戚与共，形成了优良的精神传统和鲜明的办学特色。一代代清华人“自强不息、厚德载物”，涌现出众多学术大师、兴业英才和治国栋梁，为中国社会进步和世界文明发展作出了重要贡献。特别是近年来，在国家的大力支持下，学校致力于世界一流大学建设，积极探索中国特色的“大学之道”，各项事业不断取得新的进展，正在跻身世界一流大学的行列。
大学之道，育人为本。一年来，以“清华新百年人才培养的使命与战略”为主题的第23次教育工作讨论会顺利举行，全校师生就推动办学优势转化、培养拔尖创新人才进一步取得共识。“清华学堂人才培养计划”以及多项教育教学改革措施相继实施。招生工作大力推进多元评价、兼顾拔尖与公平，生源质量进一步提高。就业毕业生中，超过80%选择到国家重要行业和领域建功立业。外国留学生规模不断扩大，结构进一步优化，外国研究生在学规模居全国高校首位。
大学之道，学术为魂。一年来，我校积极面向国际学术研究前沿和国家重大战略需求开展高

タイ語:


iGoogle - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.chula.ac.th/
Google
จามจุรี 38
ขอเชิญชาวจุฬาฯ ร่วมนับถอยหลังสู่การเป็นเจ้าภาพ "จามจุรีเกมส์"
กีฬามหาวิทยาลัยแห่งประเทศไทย ครั้งที่ 38 วันที่ 15-22 มกราคม 2554
Chulalongkorn University
จุฬาลงกรณ์มหาวิทยาลัย
เสาหลักของแผ่นดิน
มหาวิทยาลัยอันดับหนึ่งของประเทศไทย
ศูนย์ประสานงานสื่อมวลชน
จุฬาฯ เปิดศูนย์ประสานงานสื่อมวลชน (Press Center)
การแข่งขันกีฬามหาวิทยาลัยฯ ครั้งที่ ๓๘ "จามจุรีเกมส์"
การรับสมัคร
คณะ สำนักวิชา และสถาบัน
ผู้มาเยือน
แนะนำจุฬา
สื่อสารองค์กร
จุฬากับนานาชาติ
การค้นคว้าวิจัย
กิจการและการบริการสังคม

ペルシャ語:

(Isfahan university of technology, Iran)

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　にsついて、
  それを分割した単語列　$p(\mathbf{w}|s)$確率

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

  を最大にする　$\hat{\mathbf{w}}$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

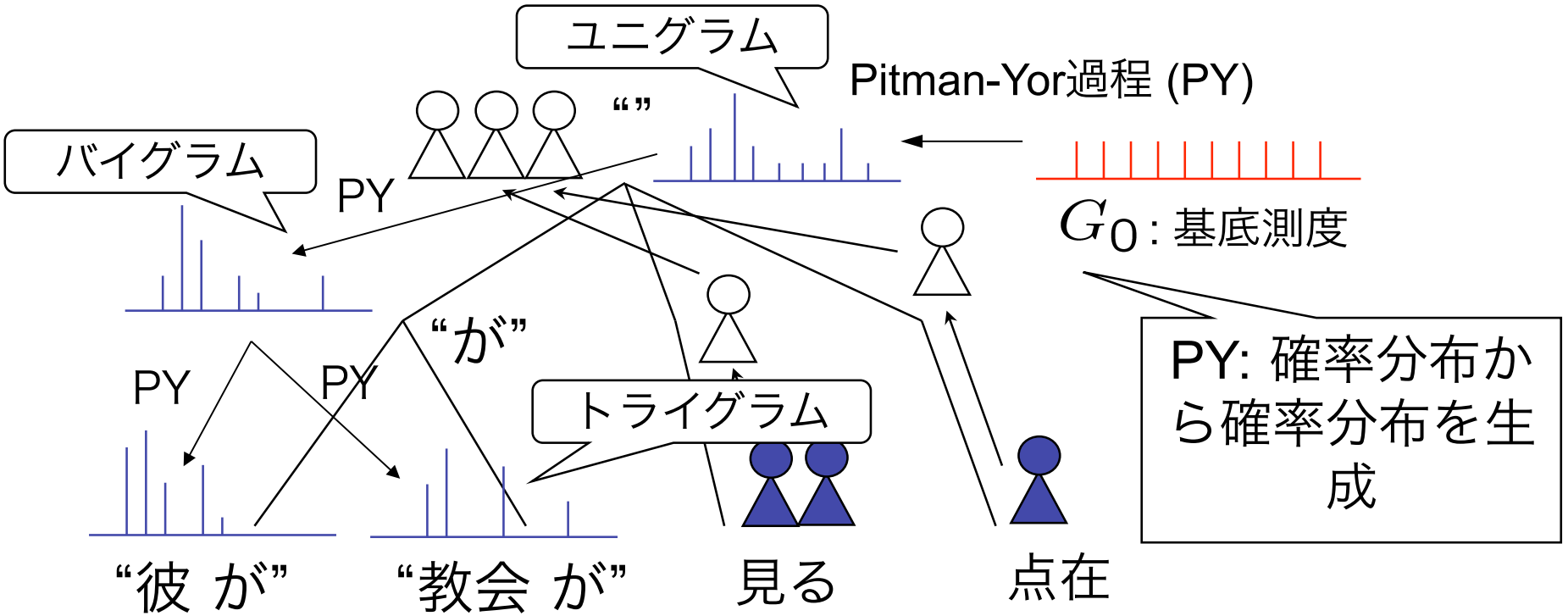
p(今日 は もう 見た)
= p(今日|^)・p(は|今日)・p(もう|は)・p(見た|もう)

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

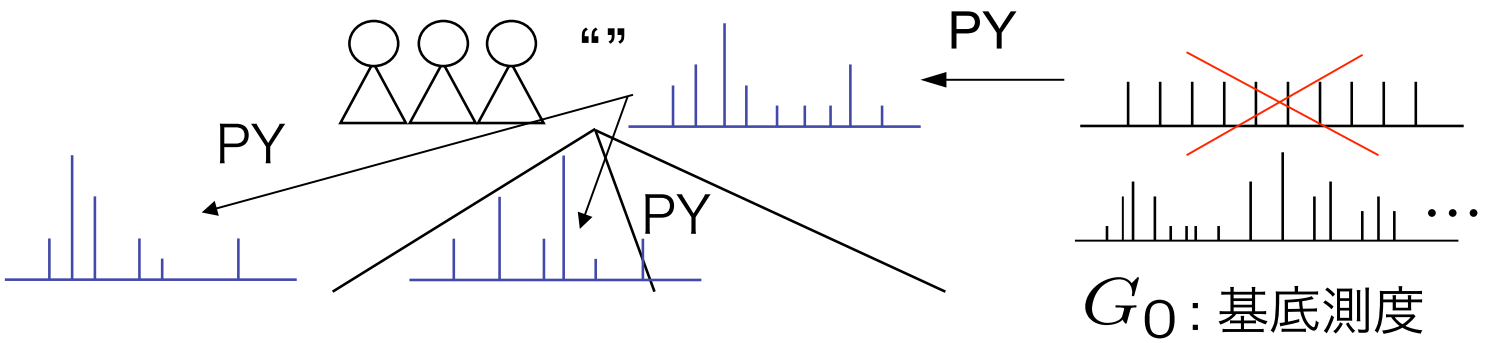
➡ 階層ベイズモデル: 階層Pitman-Yor過程言語モデル (HPYLM) (Teh 2006; Goldwater+ 2005)
- Pitman-Yor過程: ディリクレ過程 (GEM分布) の一般化

準備: HPYLM n-gram



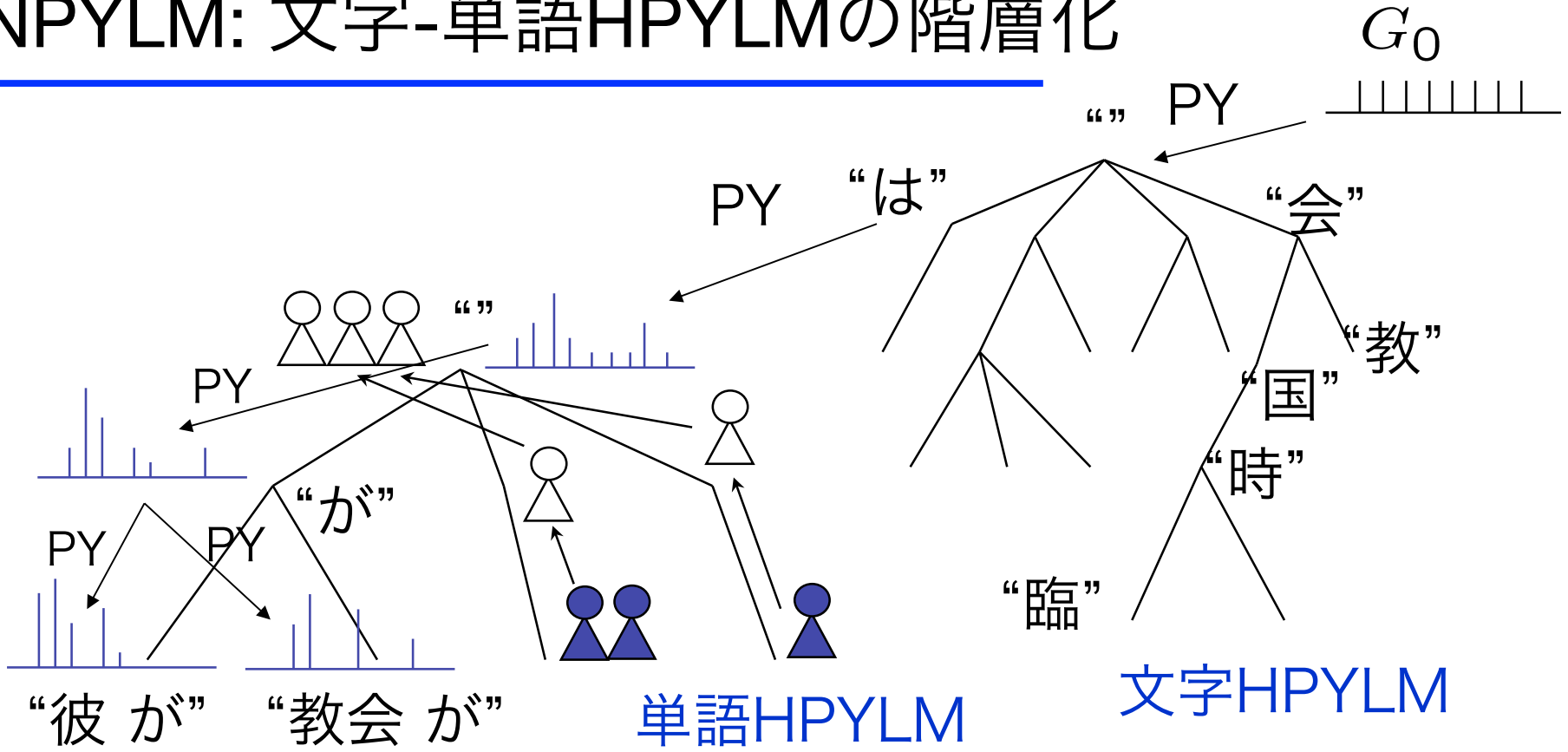
- カウントが0でも、より低いオーダーのMarkovモデルを用いて階層ベイズでスムージング
  - 注目している単語がそもそも存在しなかったら?

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

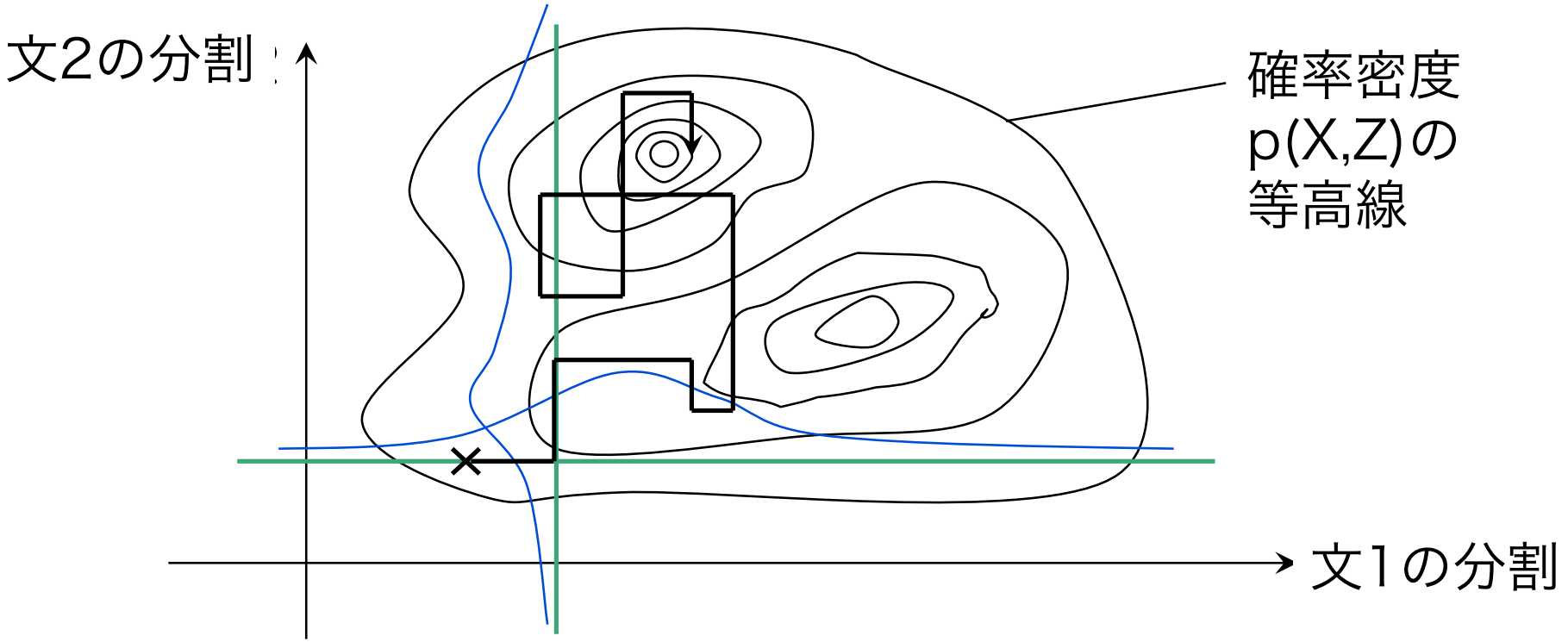
- データ: $\mathbf{X} = \{s_1, s_2, \cdots, s_X\}$ (文の集合)
  - 文: $s = c_1 c_2 \cdots c_N$ (文字列)
  - 隠れ変数: $\mathbf{z} = z_1 z_2 \cdots z_N$ ($z_i = 1$のとき単語境界)
    - 隠れ変数の組み合わせは指数的に爆発
- 文がそれぞれ独立だと仮定すると、

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

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

  - 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

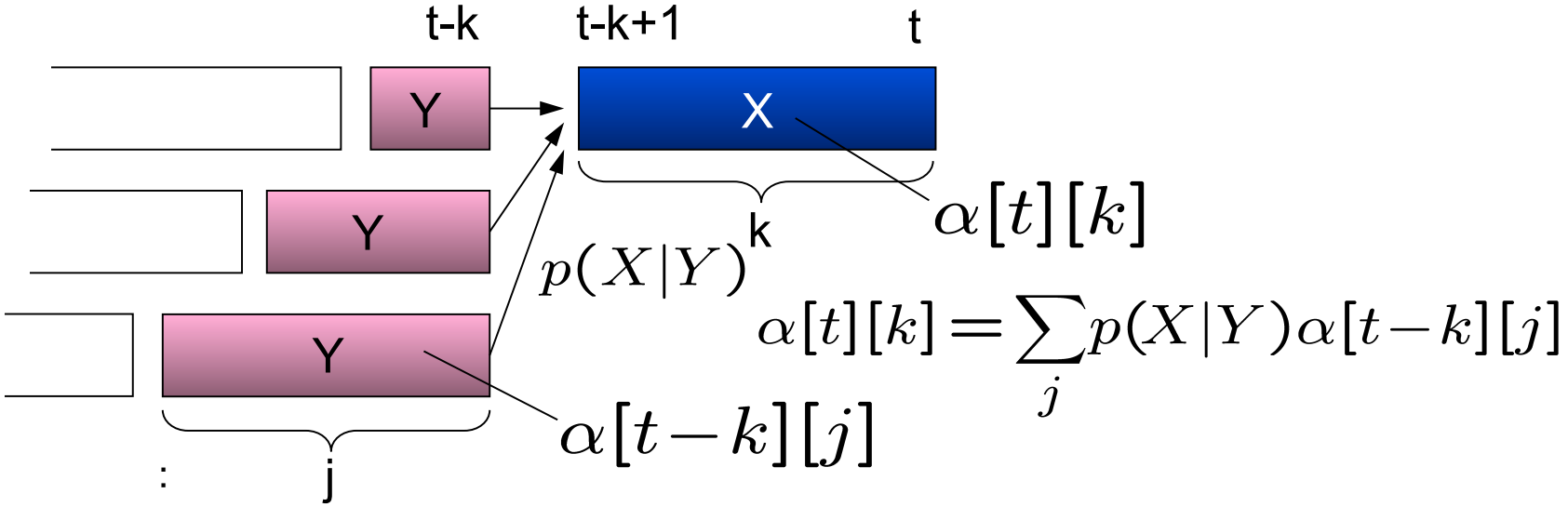
Gibbs Samplingと単語分割

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 。

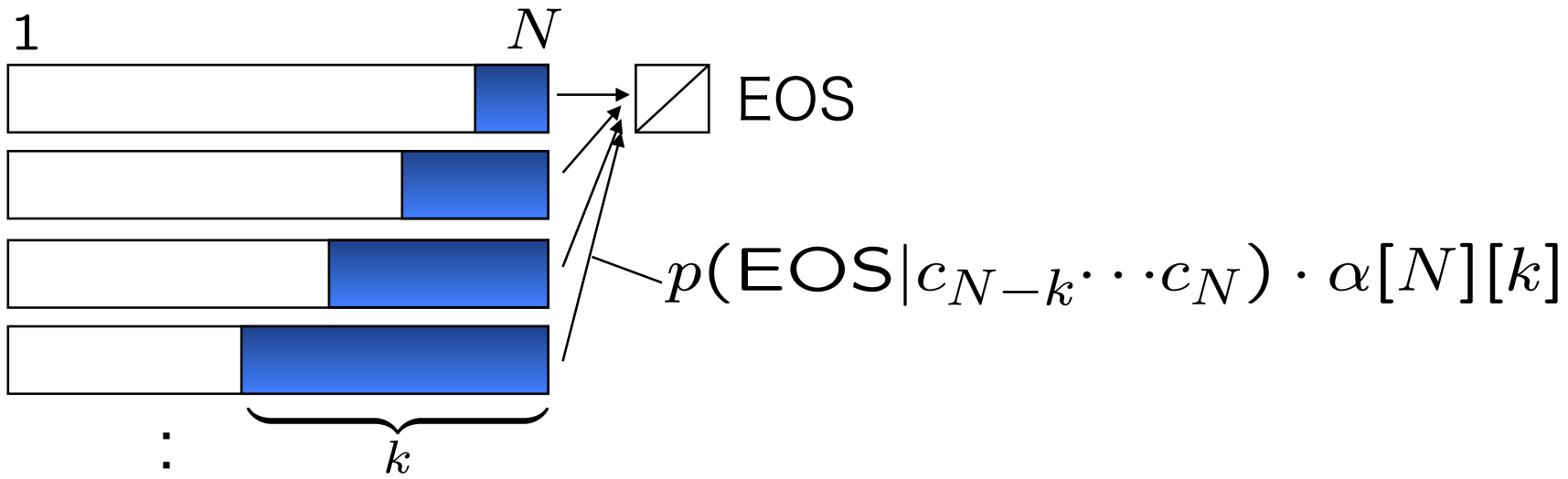
- ギブスサンプリングを繰り返すごとに、単語分割とそれに基づく言語モデルを交互に改善していく。

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

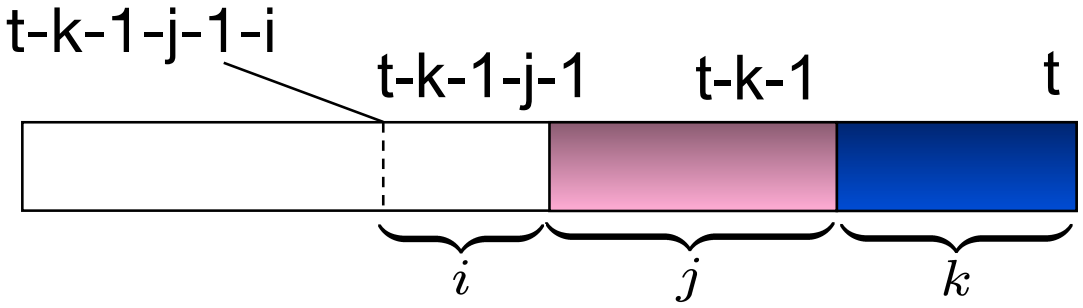


動的計画法によるデコード



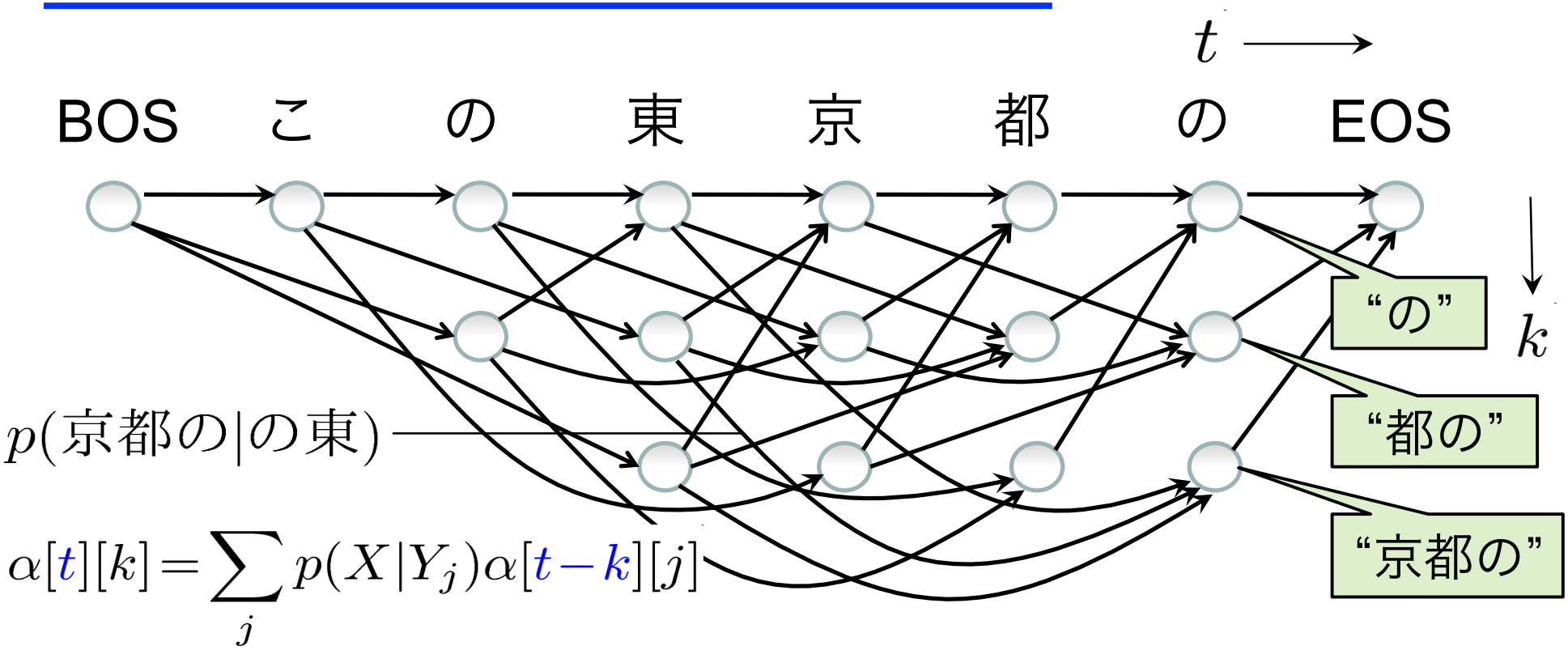
- $\alpha[N][k]$=文字列の最後のk文字が単語となる文字列確率なので、EOSに接続する確率に従って後ろからkをサンプル
- $c_{N-k}\cdots c_N$が最後の単語だとわかったので、$\alpha[N-k-1][k']$を使ってもう一つ前の単語をサンプル
- 以下文頭まで繰り返す

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)

NPYLM as a Semi-Markov model



- Semi-Markov HMM (Murphy 02, Ostendorf 96)の教師なし学習+MCMC法
- 状態遷移確率(nグラム)を超精密にスムージング

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 ・ ＳＶ シリーズ ５ 車種 を １２ 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 ８８ 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 二 トン 。

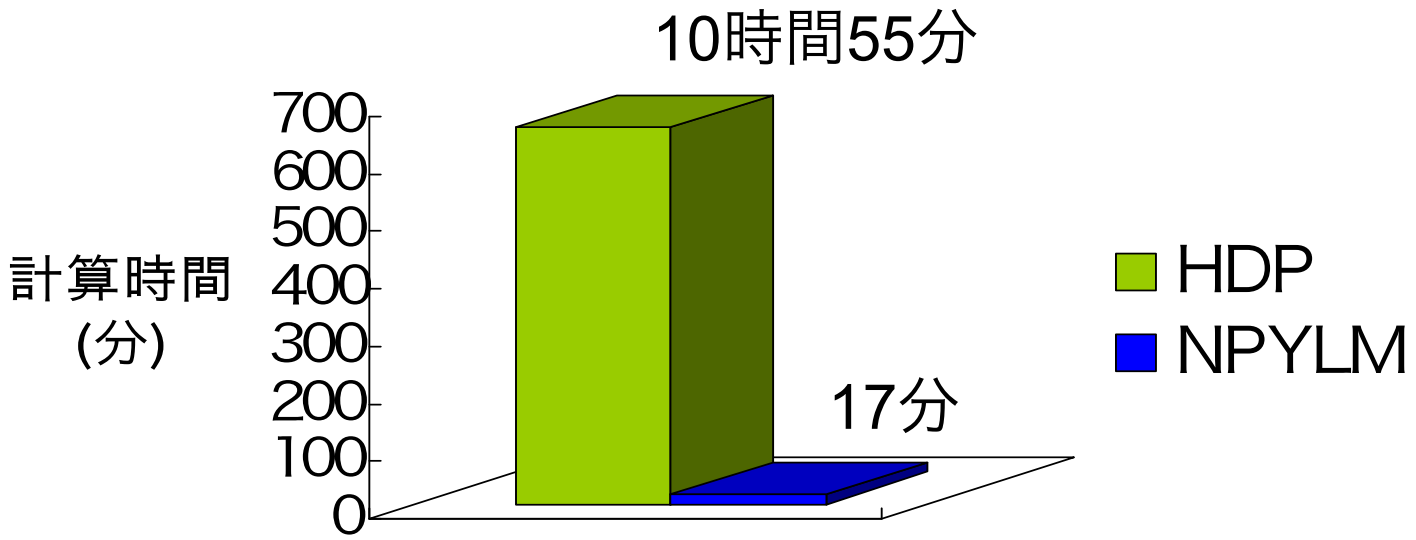
ビタビアルゴリズムで効率的に計算可能
(先行研究では不可能)

“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…

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."

وقد استغرق اعداد ه خمسة اعوام . و قال ت نادي ليي تومسون التي " تاريخي

“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 …

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

本研究のまとめ

- 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

全体のまとめ

- ノンパラメトリック・ベイズ法
  … 複雑なデータから、真に本質的な構造を取り出すための統計モデル
  - モデル選択や頻度での足切りと異なる精密なモデル
  - 潜在パラメータ数の組み合わせ爆発
- この他にも、非常に高度なモデルが存在
- 様々な分野に適用が進んでいる
  - 自然言語処理 (文法学習, 統計的機械翻訳, …)
  - 画像処理 (画像分割, 画像認識, …)
  - 機械学習全般 (IRM, Mondrian process, …)

おわり

ご清聴ありがとうございました。

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