統計的自然言語処理におけるMCMC法

持橋大地
NTTコミュニケーション科学基礎研究所
daichi@cslab.kecl.ntt.co.jp

2010-2-21(日), 統計数理研究所

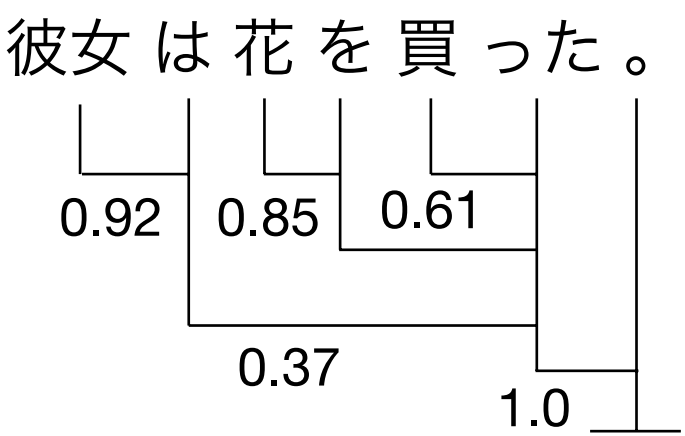
“The Gods may throw a dice..”
--- ABBA `The winner takes it all’

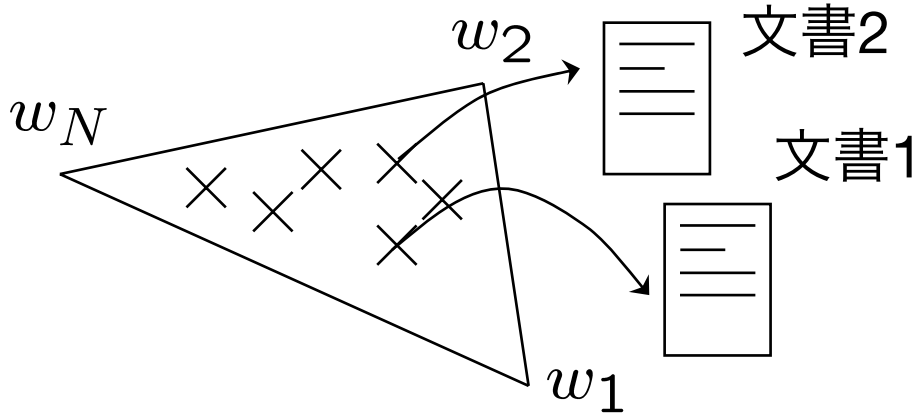
自己紹介

- NTTコミュニケーション科学基礎研究所
  リサーチアソシエイト (PD, 来年度からRS=上級研究員)
  - 京阪奈学研都市：京都から近鉄35分＋バス15分
  - NTT持株本社
  - NTT研究所の中でも、基礎研究に特化、
    超少数精鋭
    - かなり大きな建物に研究員は2ケタ
- 研究分野：統計的自然言語処理

自然言語処理とは

- 「計算言語学」ともいわれる
  - 大量のテキストデータの統計的な分析に基づく
    - 形態素解析 (単語分割, 品詞付与)
    - 構文解析・係り受け解析
    - 統計的意味解析
    - 文書の統計モデルと情報検索 etc, etc …



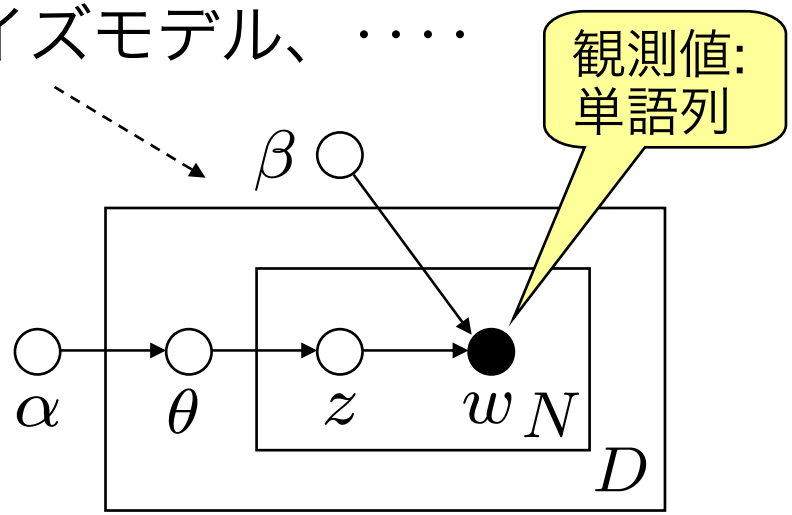


統計的自然言語処理

- 1990年代後半～からパラダイムシフト
  - 統計的機械学習の一部として重要な位置
- 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、・・・・

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

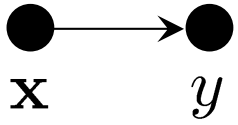
ある単語xの品詞
が形容詞である確率



自然言語処理でのモデル化と学習

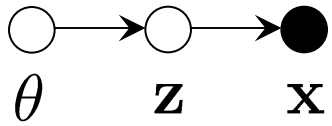
- 教師あり学習と教師なし学習
  - 教師あり学習、分類学習
    - 対数線形モデル

$$\log p(y|\mathbf{x}, \Lambda) \propto \sum_i \lambda_i f_i(y, \mathbf{x})$$



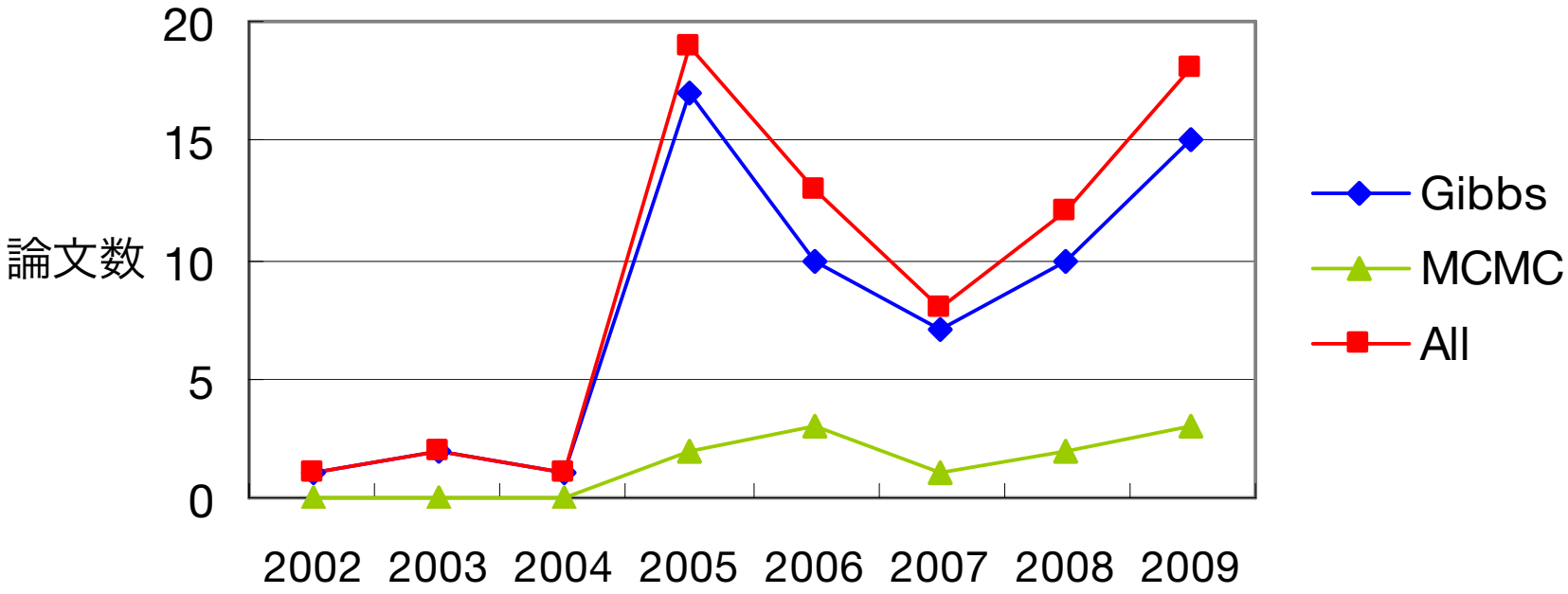
  - 教師なし学習 (自己組織化)
    - 生成モデル、階層モデル

$$p(\mathbf{x}|\theta) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\theta)d\mathbf{z}$$



    - 対数線形モデル、ボルツマンマシン (一部)
- 特徴：離散、超高次元、(超)大規模学習
  - 数万～数100万次元の離散分布、
  - 数千万語～数億語のデータ

統計的自然言語処理とMCMC



- ACL: Association of Computational Linguistics
  計算言語学/自然言語処理の分野のトップ国際会議の論文中の検索数

Why MCMC lately?

- 従来の自然言語処理の方法：
  最尤推定、EMアルゴリズム

モデルの複雑化、精緻化
(大域依存性、階層モデル)

- 最近のアプローチ：
  ベイズ推定、変分ベイズEMアルゴリズム、
  MCMC法
  (SMC, EP+ など… not yet explored)
  - 特にMCMCは、局所解に陥らない＆実装が簡単

MCMCが用いられている具体的な問題の例

- ✓教師なし品詞解析
  - Goldwater+ (2007), Johnson+ (2007), Gao+(2008) など
- 構文解析、係り受け解析
  - Johnson+ (2007)、中川 (2007)など
- ✓潜在的意味解析ー Latent Dirichlet Allocation
  - Blei (2001), Griffiths+(2006)
- ✓教師なし単語分割
  - Goldwater+ (2006), 持橋 (2009)
- ✓潜在語言語モデル
  - Deschacht+ (2009)

他に、統計的機械翻訳など
(DeNero 08)(Cohn+09)etc..

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

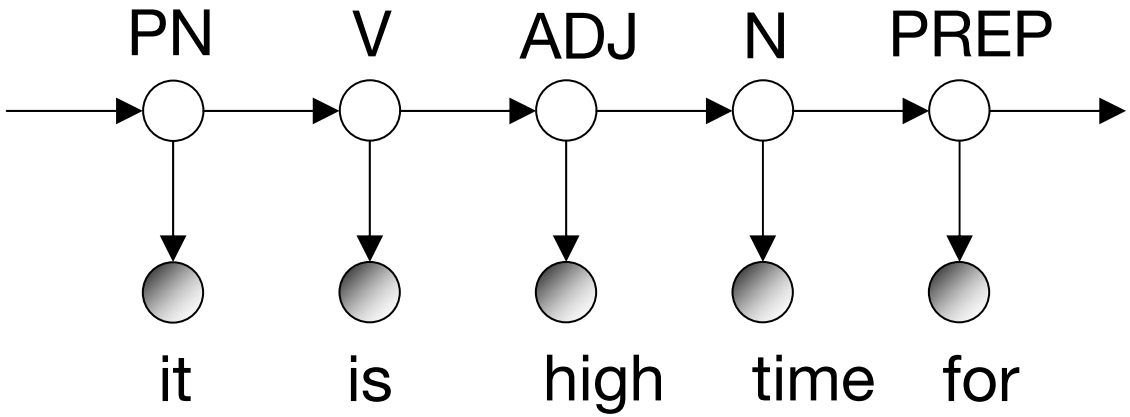
  たなびく 雲 の 合間 から 漏れ出る 静かな 月 。
  動詞 名詞 助詞 名詞 助詞 動詞 形容動詞 名詞

  After procrastination, it is the time for prime-ministerial leadership .
  PREP N PN V 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)

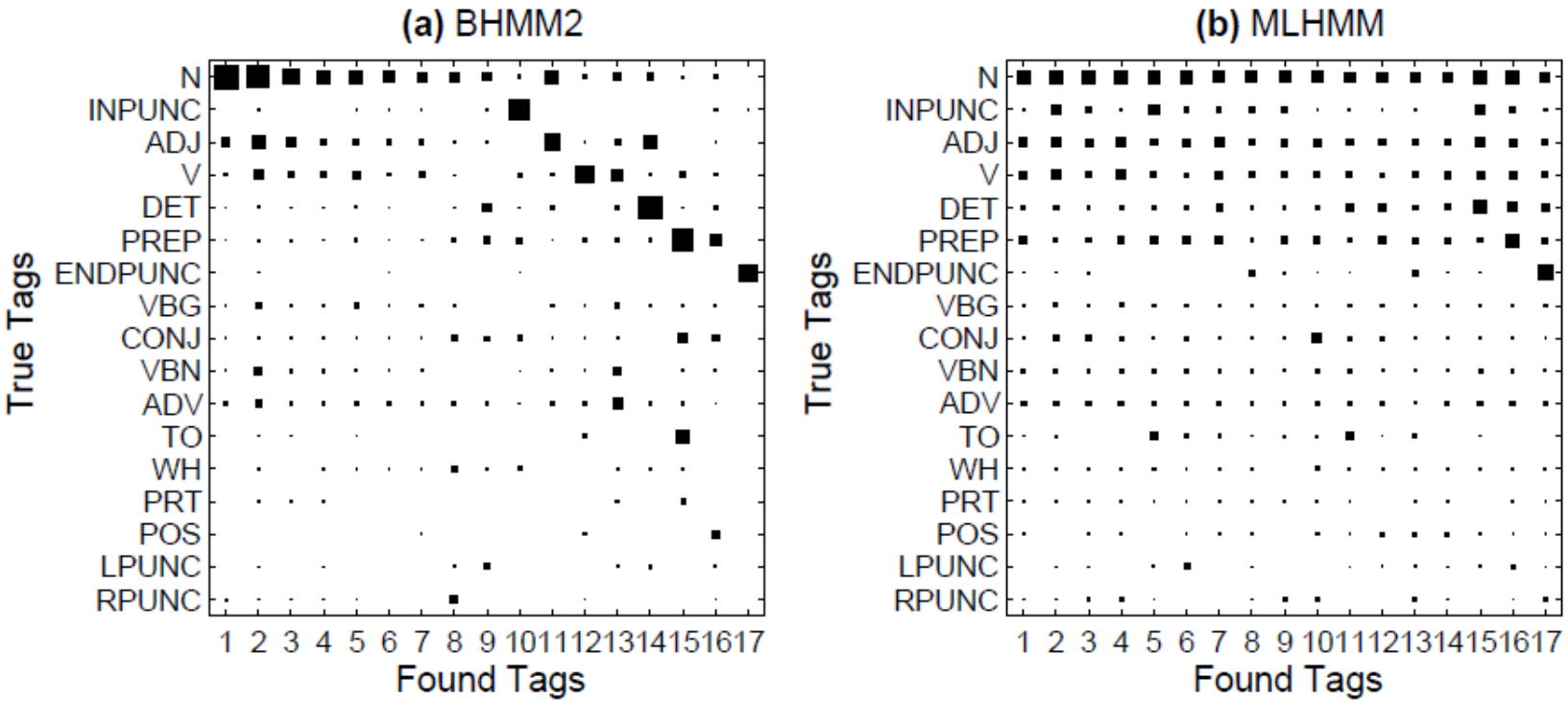
- 基本的なアプローチ: HMM (隠れマルコフモデル)を使おう
  - 「隠れ状態」が品詞に対応する(はず)



- 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

$$\mathrm{P}(y_i|\mathbf{x},\mathbf{y}_{-i},\alpha) \propto \left(\frac{n_{x_i,y_i}+\alpha_x}{n_{y_i}+m\alpha_x}\right)\left(\frac{n_{y_i,y_{i-1}}+\alpha_y}{n_{y_{i-1}}+s\alpha_y}\right)\left(\frac{n_{y_{i+1},y_i}+\mathrm{I}(y_{i-1}=y_i=y_{i+1})+\alpha_y}{n_{y_i}+\mathrm{I}(y_{i-1}=y_i)}\right)$$

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)

確率的潜在意味解析 (1)

- 言語や文書に隠れた「意味」を知りたい

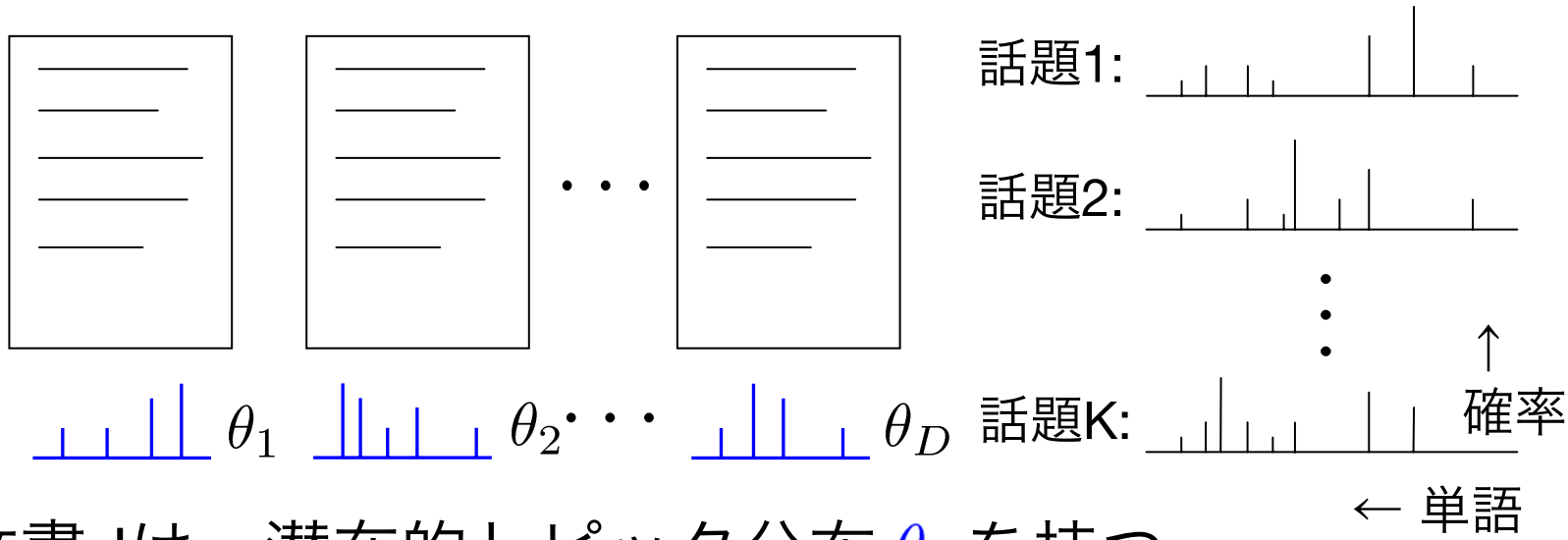
スエズ運河の整備計画は五年目を迎え、国境では通行税の‥	スキー競技の選考会が行われた二十日未明から嵐に変わり‥
→ 「国際」 「開発」	→ 「スポーツ」 「天気」

- 文書にはどんな話題が潜在的に隠れているか?
  - この単語はどんな話題に関連しているか?
  - 複数の話題の混じった文書（が普通）
  - どんな話題がそもそも存在しているか?
- 文書データは、ただ単語が並んでいるだけ

確率的潜在意味解析 (2)

遺伝学では，Admixtureと呼ばれる

- Latent Dirichlet Allocation (Blei+ 2001)
  - Probabilistic LSI (Hofmann 1999)のベイズ化
  - 遺伝学分野でのPritchard(2000)と基本的に同じモデル



- 各文書dは、潜在的トピック分布 θ_d を持つ
  - θ_d からn番目の単語の潜在トピック z_{dn} をサンプル
  - z_{dn} から単語 w_{dn} を生成することで、文書が生成

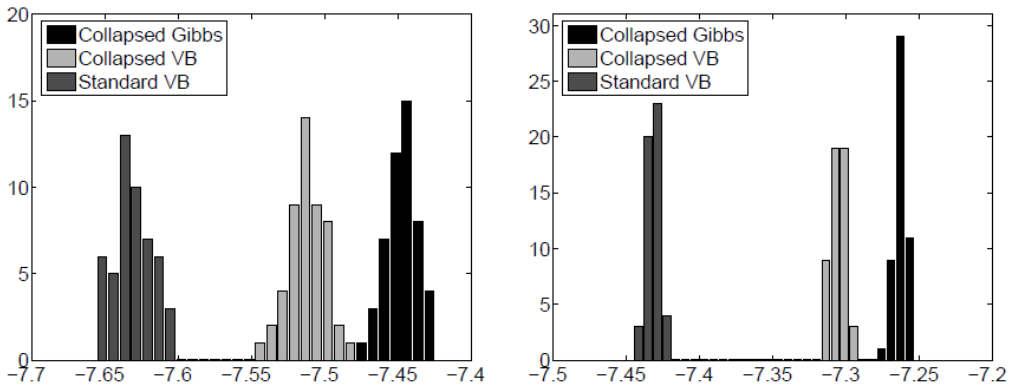
確率的潜在意味解析 (3)

- LDAのベイズ学習
  - Gibbsが正確で易しい

$$p(z_{dn}=k|w_{dn}=w) \propto \frac{n^{-}(w,k)+\beta}{\sum_w n^{-}(w,k)+\beta} \cdot \frac{n^{-}(d,k)+\alpha_k}{\sum_k n^{-}(d,k)+\alpha_k}$$

$p(w|k)$ $p(k|d)$

  - Pritchard (2000)では, θ も積分消去せずにサンプリング
  - 上の式は、機械学習ではCollapsed Gibbsと呼ばれる
    - 統計用語では, Rao-Blackwellized Gibbs



VB(変分ベイズ),
Collapsed VBと比べ,
Gibbsが最も高い性能

確率的潜在意味解析 (4)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

- 各単語に、その単語を生成した潜在的なトピックが学習される
  - 単語の持つ大まかな「意味」が教師なしでわかる
  - 文書の持つトピック分布は、基本的にそれらの和

潜在語言語モデル (1)

- 自然言語処理の大きな問題：超高次元＆離散
  - 単語が(たとえ意味的関連性があっても)まったく別の次元として扱われる

"Most of the confidences were unsought—frequently I have feigned sleep, preoccupation, or a hostile levity when I realized by some unmistakable sign that an intimate revelation was quivering on.." (from "*The Great Gatsby*")

- 観測された単語の裏には何もないのか?

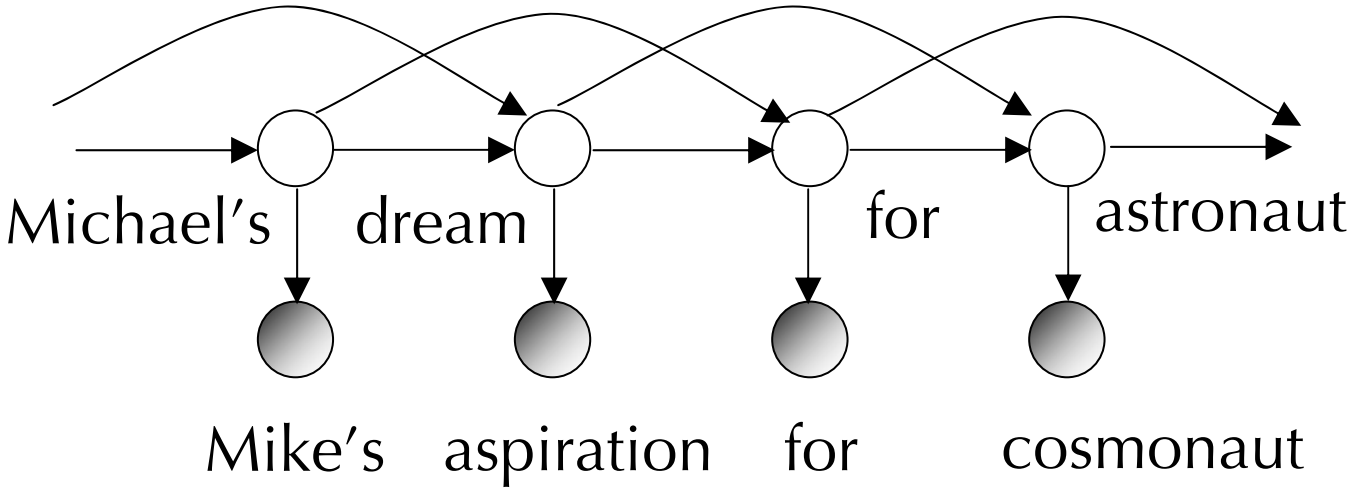
Michael thought he ought to be a cosmonaut.

Andrew, Bob, Susan　should, have to　astronaut, aeronaut

  - ソシュールの「範列」(paradigme)の考え方!

潜在語言語モデル (2)

- The Latent Words Language Model (Deschacht+ 09)



  - 観測された各単語は、対応する潜在語から生成された
  - 一種の隠れMarkovモデル (ただし超高次元)
- 単語ー単語の翻訳確率 p(w|w') が存在, もちろん未知
  - データの尤度を最大化する潜在語と翻訳確率を学習

潜在語言語モデル (3)

- 学習‥ギブスサンプラで、各単語の潜在語を次々とサンプル

$$p(h_t|w_t) \propto p(w_t|h_t) \times p(h_t|h_{t-1}, h_{t-2}) \cdot p(h_{t+1}|h_t, h_{t-1}) \cdot p(h_{t+2}|h_{t+1}, h_t)$$

  - $p(w_t|h_t)$ は(事前)翻訳確率で、データ全体での置換回数+Dirichlet priorから求まる
  - $p(h_i|h_{i-1}, h_{i-2})$ は潜在語のnグラム確率(ここでは3グラム)で、サンプルされた潜在語によって動的に変化
  - 潜在語の初期値は観測値と同じ

C++での実装

```
for (j = 0; j < iter; j++) {   // repeat for many sweeps
  random_shuffle (words.begin(), words.end(), irand);   // visit randomly

  for (it = words.begin(); it != words.end(); it++) {
    for (n = order; n >= 0; n--)
      lm->remove_customer (hidden + n);                  // remove from LM
    if (j > 0) table->sub_count (*hidden, *observed); // remove from table
    *hidden = draw_gibbs (hidden, observed, table); // sample latent word
    table->add_count (*hidden, *observed);             // add to table
    for (n = 0; n <= order; n++)                       // add to LM
      lm->add_customer (hidden + n);
  }
}
```

- サンプルする単語数は通常、数百万～数億単語なので効率的な実装が不可欠

京大コーパス (毎日新聞)の学習例

Original:
しかも 、 政 ・ 官 ・ 業 の 鉄 の トライアングル は 、 これ ら の 業界 の 中 で 増殖 して いった 。
今年 は 「 規制 緩和 」 の 合唱 が 始まって 三 年 目 に なる 。
細川 政権 下 、 平岩 リポート が 「 経済 的 規制 は 原則 自由 」 と いう 提言 を 発表 した の は 、一昨年 十一 月 であった 。
一 つ の 合唱 が やがて 二 つ 、 三 つ と 輪 を 広げて いった 。
小倉 氏 が 投じた 一石 は やがて 巨大な 新 産業 を 生む が 、 同時に 各 業界 を 規制 する 「 事業 法 」 の 在り 方 を 見直す 論議 に 火 を つけた 。

Sampled Latent Words:
しかも 、 政 ・ 官 ・ 業 の 鉄 の 点 は 、 これ まで の 政策 の 中 で 発生 して いる 。
今年 は 「 規制 緩和 」 の 気持ち が 、 二 年 目 に なる 正直な
細川 内閣 間 、 玉虫色 対応 が 「 経済 的な 転換 は 「 再生 」 と いう 見解 を 強調 した の は 、 昨年 九 月 の 主柱
一 つ の 気持ち が 、 一 つ の 三 つ の 場 を 広げて いる 。
海部 氏 が つなげ 再建 は 、 ＪＲ 新 産業 を 高める が 、 今 の 利益 を 廃止 する 「 事業 法 」 の 在り 方 を 狙った 行政 の 意識 を 開く 。

Austen Novel (“Emma”)

Original:
emma woodhouse handsome clever and rich with a comfortable home and happy disposition seemed to unite some of the best blessings of existence and had lived nearly twenty one years in the world with very little to distress or vex her
she was the youngest of the two daughters of a most affectionate indulgent father and had in consequence of her sister's marriage been mistress of his house from a very early period…

Sampled Latent Words:
emma remained in love and their for a hereabouts home and artifice ours were to persuade some of the best exciting of him she had already moves but one draper in the least and how little to her or handle her he was the children of the two daughters of a most delightful stronger myself and and in scruples of his father's having been lateness of his wife and a very great difference…

Austen latent word translations

would	manners	0.003	1
would	might	0.075	25
would	must	0.087	29
would	need	0.030	10
would	never	0.006	2
would	often	0.015	5
would	or	0.003	1
would	possibly	0.009	3
would	should	0.249	83
would	spirits	0.006	2
would	triumph	0.003	1
would	would	0.477	159

randalls	dinner	0.0489	3
randalls	home	0.538	33
randalls	least	0.1957	12
randalls	various	0.0489	3

opportunity	arrangement	0.0566	2
opportunity	attachment	0.0849	3
opportunity	evening	0.1698	6
opportunity	inducement	0.0566	2
opportunity	interval	0.0566	2
opportunity	opportunity	0.1132	4

father	daughter	0.0670	9
father	friends	0.0372	5
father	grandmama	0.0298	4
father	having	0.0149	2
father	head	0.0447	6
father	inferior	0.0149	2
father	master	0.0298	4
father	memory	0.0298	4
father	mother	0.2680	36

abbey	abbey	0.7240	14

LWLM論文での例

no	**plans**	**to**	**dissolve**	**the**	**house**
no 0.93	plans 0.83	to 0.96	dissolve 0.54	the 0.99	*house* 0.84
great 0.05	aims 0.03	towards 0.02	grant 0.42	its 0.01	*parliament* 0.08
any 0.01	offers 0.03	raises 0.01	alleviate 0.04		*generals* 0.01
then	**set**	**the**	**house**	**ablaze**	**.**
then 0.92	set 0.74	the 0.94	*house* 0.40	ablase 0.63	. 0.98
also 0.03	shot 0.12	their 0.01	*hotel* 0.13	unnaceptable 0.27	; 0.01
nonetheless 0.01	place 0.04	his 0.01	*trees* 0.02		? 0.01
her	**one-storey**	**house**	**in**	**centre**	**of**
her 0.63	own 0.12	*house* 0.70	in 0.84	centre 0.92	of 0.99
your 0.22	upper 0.06	*home* 0.05	near 0.02	routes 0.04	without 0.01
his 0.09	brick 0.02	*property* 0.01	into 0.02	opposition 0.01	from 0.01

Table 3: Probabilitistic expanded words for three phrases of the the *Reuters* corpus, containing the word "house". Note how the expansion was different for the different contexts.

- 同意語は文脈依存になっていることに注意
- 言語モデルとしての予測性能も改善
  - PPL 102.8 (K-N 4-gram) → 93.65 (LWLM 4-gram)

LWLM: 注意

- シンプルな話だが、誰も挑戦していなかった→Why?
  - 期待値を計算するEMアルゴリズムでは、事後分布の次元が各単語について数万次元 (メモリ爆発)
  - MCMCによるサンプリングが不可欠
  - 「潜在変数は低次元でないといけない」という思い込み
- 学習が遅いのでは?
  → ナイーブな方法は確かに遅い
  - しかし、Beam samplingで高速化することが原理的に可能
- 実は、自然言語処理一般に有用な興味深い研究

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, 統数研はどうですか。“
 | mecab -O wakati
やあ こんにちは , 統数研 は どう です か 。
(やあこん にち は , 統 数研 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|、|…

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

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

  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)

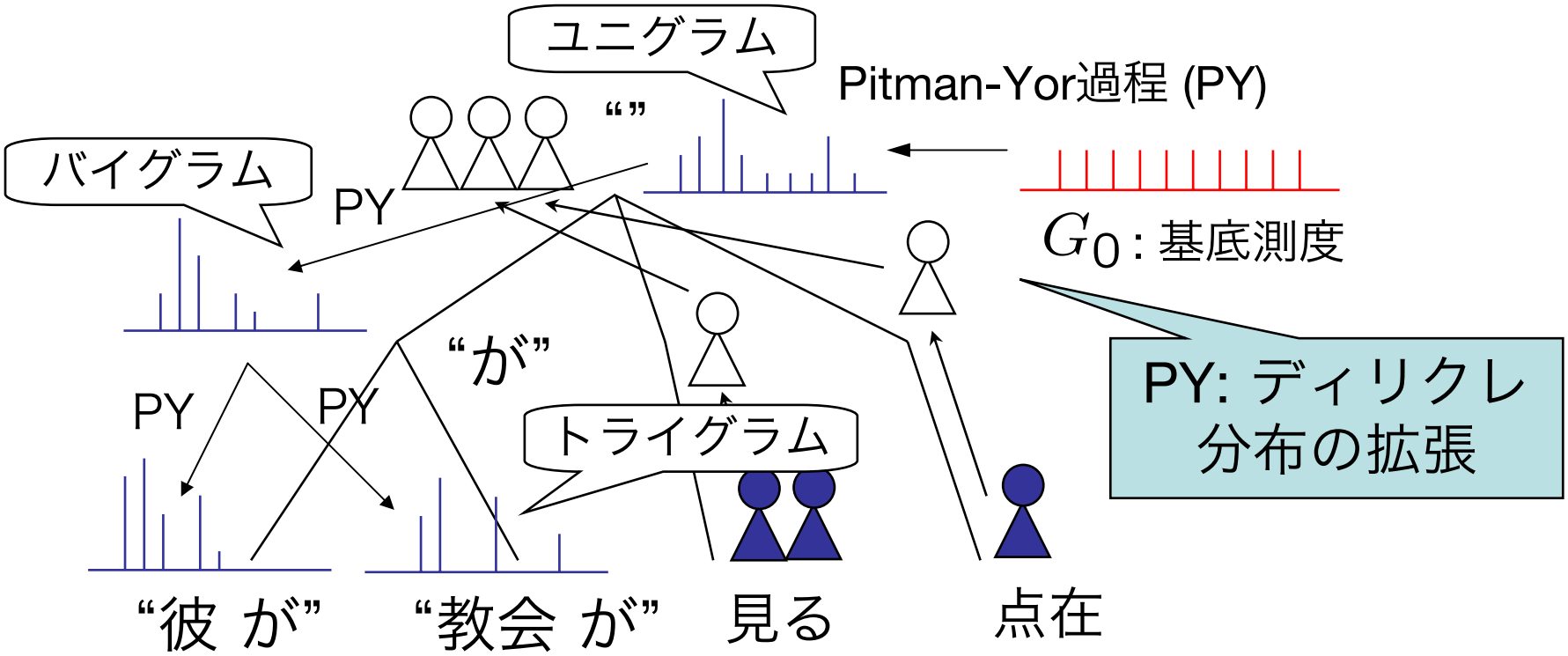
文頭を表す特殊文字

- 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)

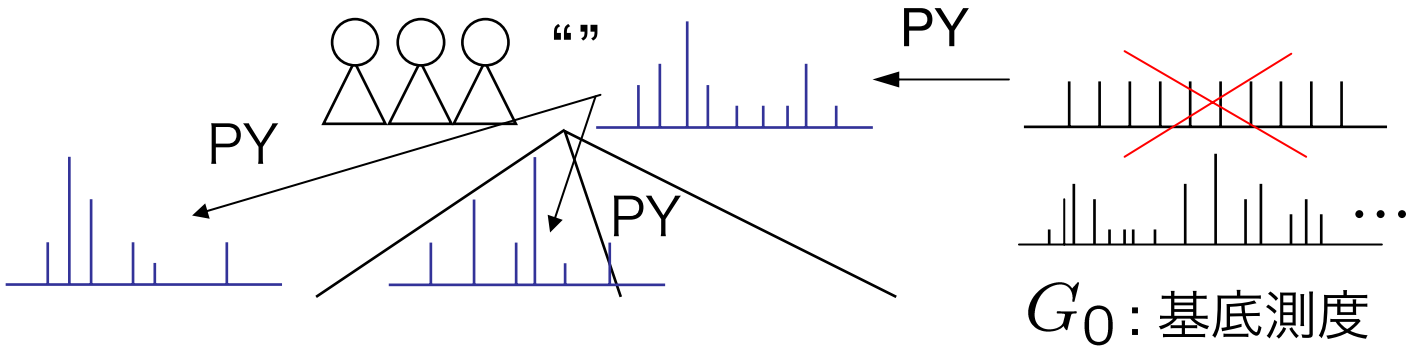
- Pitman-Yor過程: ディリクレ過程 (無限次元ディリクレ分布) の拡張

準備: HPYLM n-gram



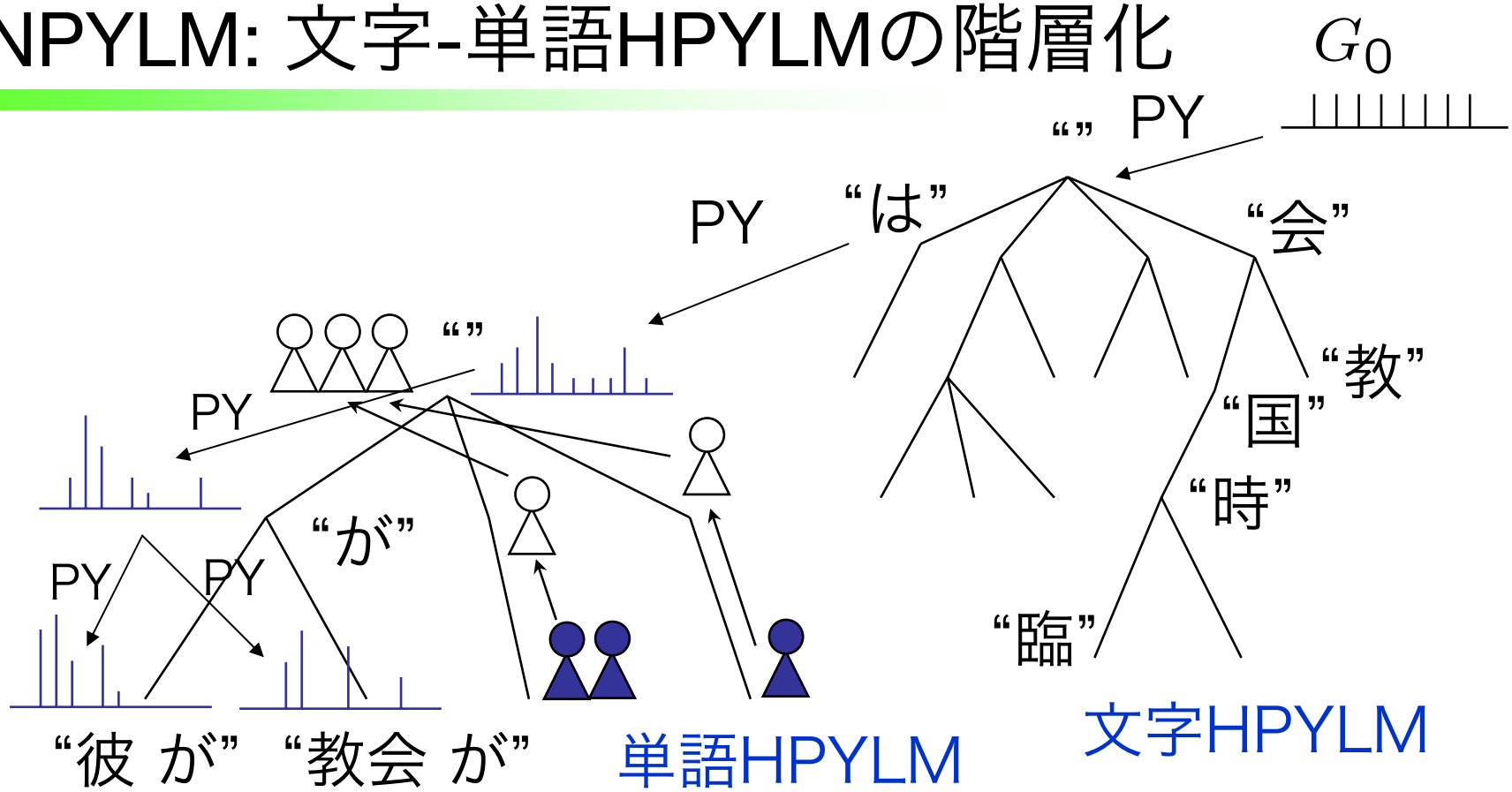
- カウントが0でも、より低いオーダーのMarkovモデルを用いて階層ベイズでスムージング
  - 注目している単語がそもそも存在しなかったら?

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

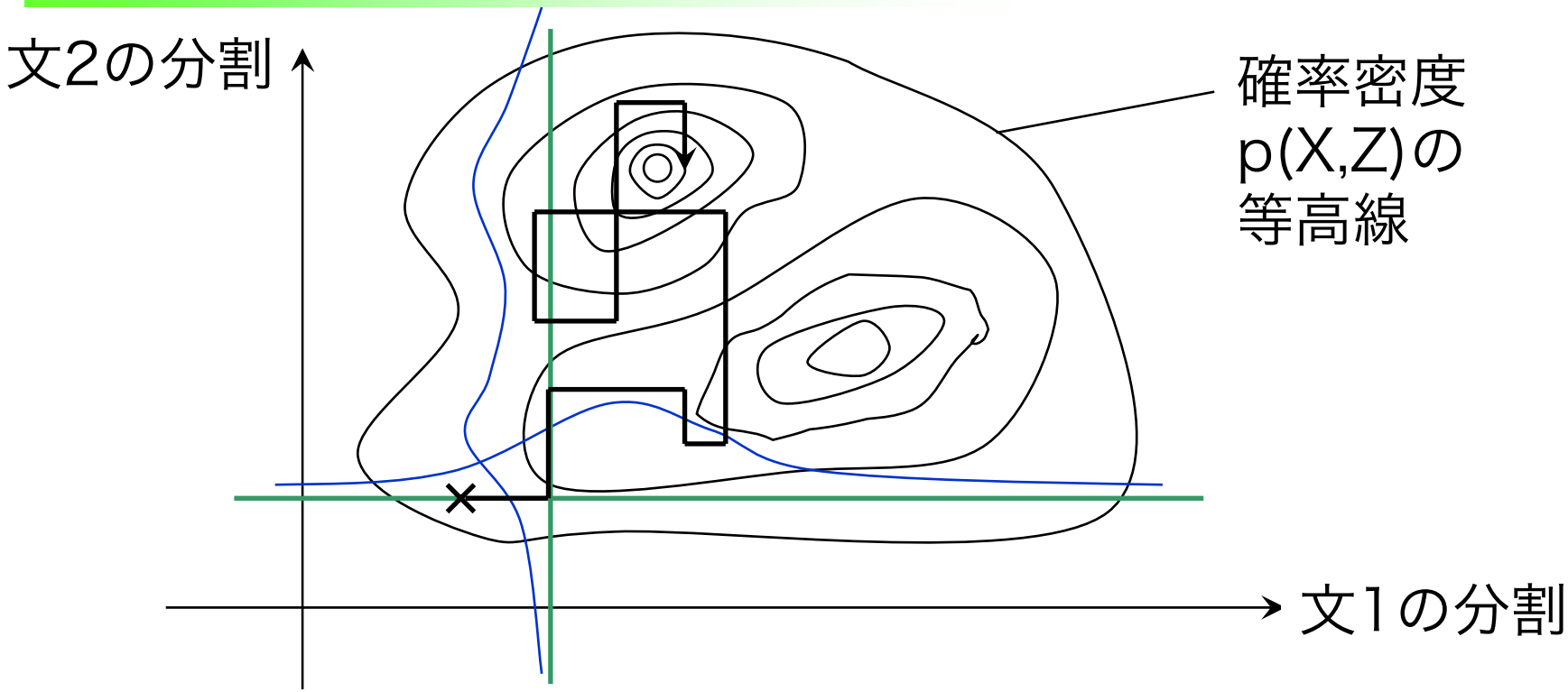
- データ: $\mathbf{X} = \{s_1, s_2, \cdots, s_X\}$ (文の集合)
  - 文: $s = c_1 c_2 \cdots c_N$ (文字列)
  - 隠れ変数: $\mathbf{z} = z_1 z_2 \cdots z_N$ ($z_i = 1$ のとき単語境界)
    - 隠れ変数の組み合わせは指数的に爆発
- 文がそれぞれ独立だと仮定すると、

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

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

  - 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

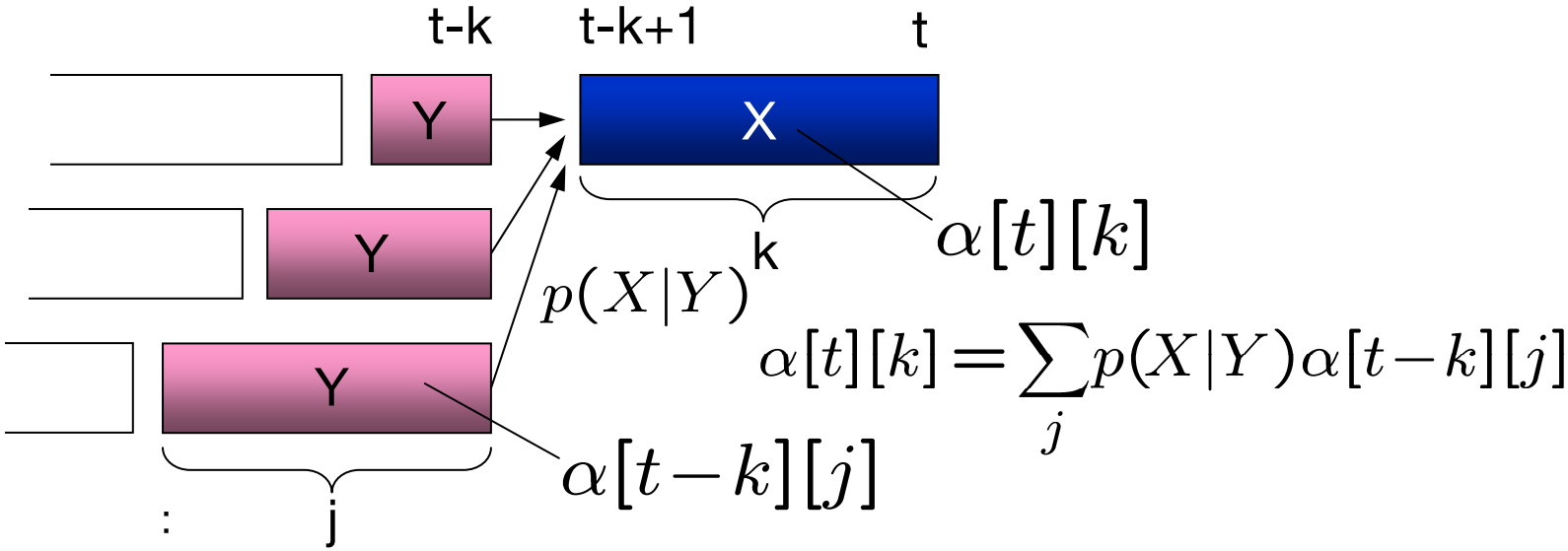
Gibbs Samplingと単語分割

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 。

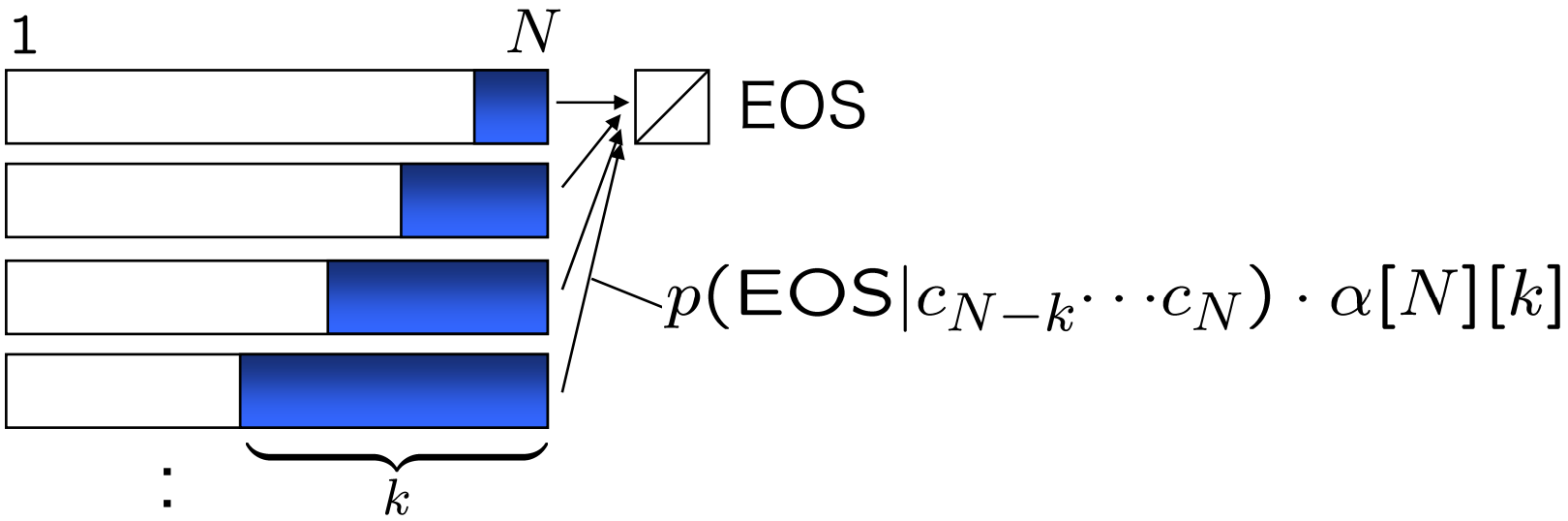
- ギブスサンプリングを繰り返すごとに、単語分割とそれに基づく言語モデルを交互に改善していく。

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

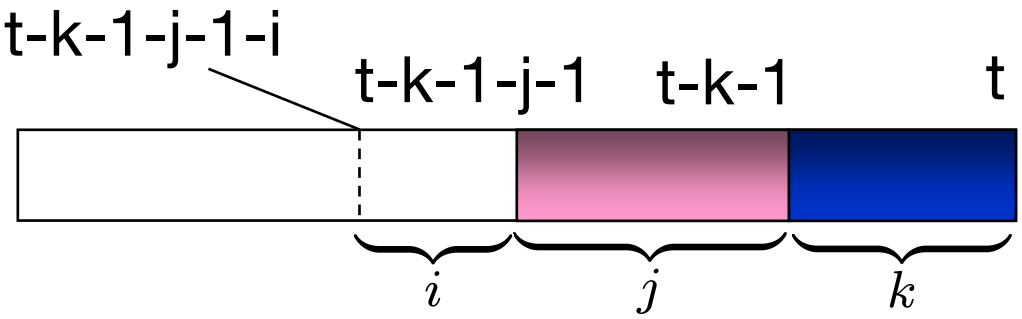


動的計画法によるデコード



- $\alpha[N][k]$=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)

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 。

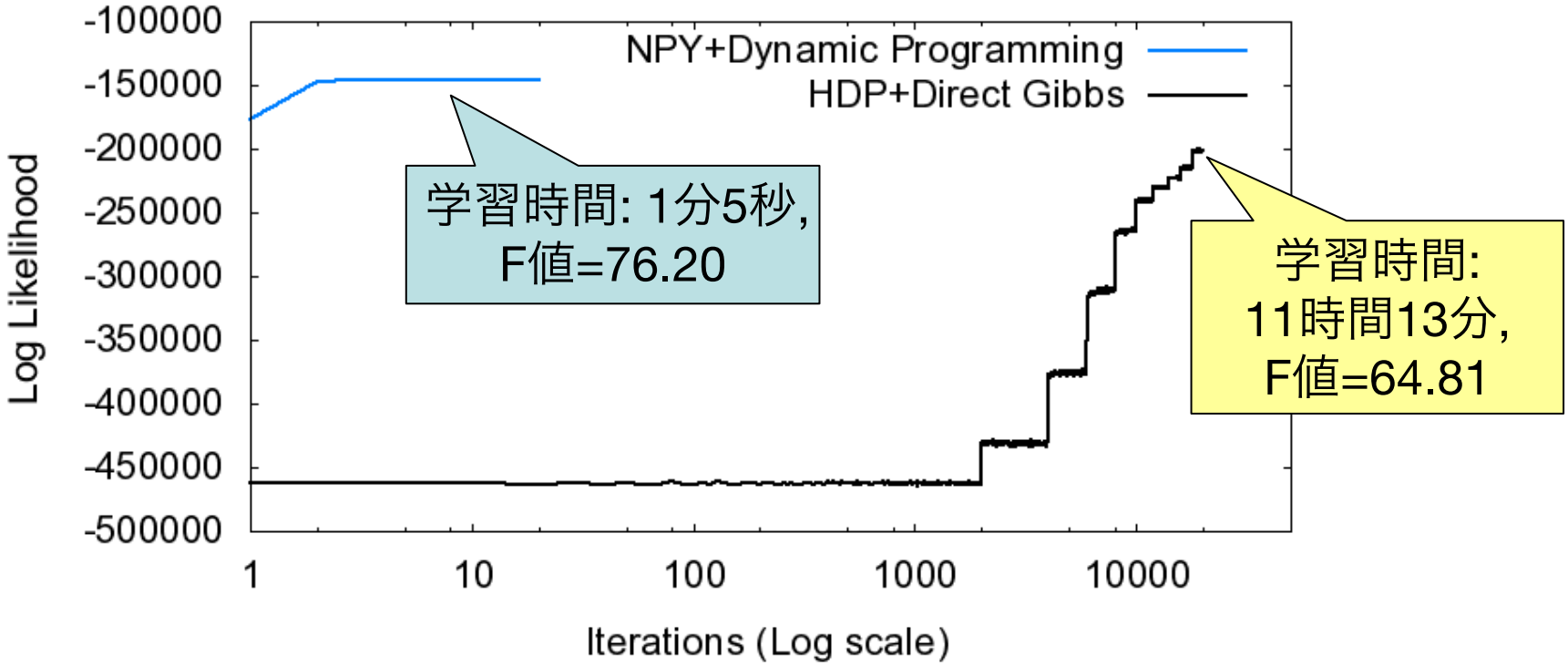
ビタビアルゴリズムで効率的に計算可能
(先行研究では不可能)

“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

計算時間と収束の比較



- 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)
- 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‥‥

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)

الفلسطينييسببتظاهرةلانصارحركةالمقاومةالاسلاميةحماس.

و اذاتحققذلكفانكيسلو فسكيكو نقدحانثلاثة جرىفيابرز ثلاثة

صحية

قائد+

الايقل

وقالتدانييلتومسونالتيكتبتالسيناريو.وقداستغرقاعدادهخمسةاعوام."تاريخي

Google translate:
“Filstinebsbptazahrplansarhrkpalmquaompalaslamiphamas.”

NPYLM

الفلسطيني بسبب تظاهرة ل انصار حركة المقاومة الاسلامية حماس .

و اذا تحقق ذلك ف ان كيسلوفسكي يكون قد حان ثلاث جرىفيابرز ثلاثة

صحية

سطينية + " .

مالايقل

وقد استغرق اعداد ه خمسةاعوام . و قال ت دان ييل تومسون التي " تاريخي

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

- MCMC法は、最近の複雑な統計的自然言語処理の学習において重要なツール
  - 組み合わせ最適化の塊、EMではすぐに局所解
- 事後分布がきわめて高次元or無限次元
  →サンプリングが不可欠な場合
  - 隠れ単語、隠れ構文木、隠れカテゴリ、…
- 非常に大規模な学習、高効率な実装が必要
  - 数千万～数億語のデータ, C++等で高速な実装
  - 大量のデータ処理のためのMCMCの並列化やそのためのモデル化も最近様々に提案されている

展望

- 現状の自然言語処理は、対数線形モデル＝最適化、ベイズモデル＝EM/MCMC と大きく二分されている
  - 人手教師データの分類性能を上げたいなら、前者が有利
- ベイズモデルでも、EM派とMCMC派が存在
  - 両方のいい所取り?

- きわめて最近提案! (Carbonetto+, NIPS 2009)
  - 解析近似とSMC、最適化の組み合わせ
- サンプリングの考えを, 狭義のMCMC法に囚われず適用していくことが今後有用
  - 超大規模データでは, 全数数え上げは無理/不必要

終わり

- ご清聴ありがとうございました。