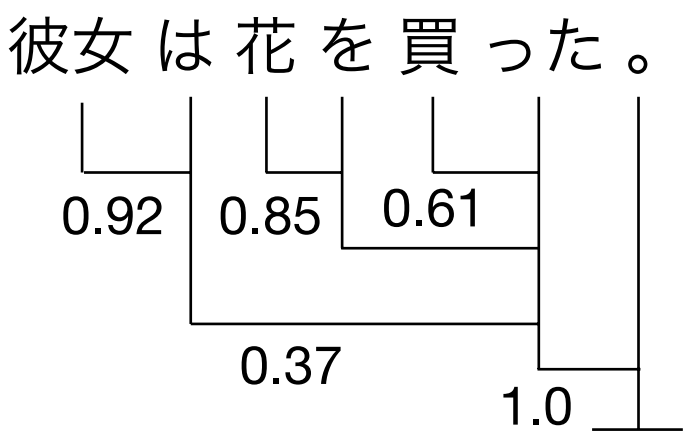
ノンパラメトリックベイズ法による教師なし形態素解析

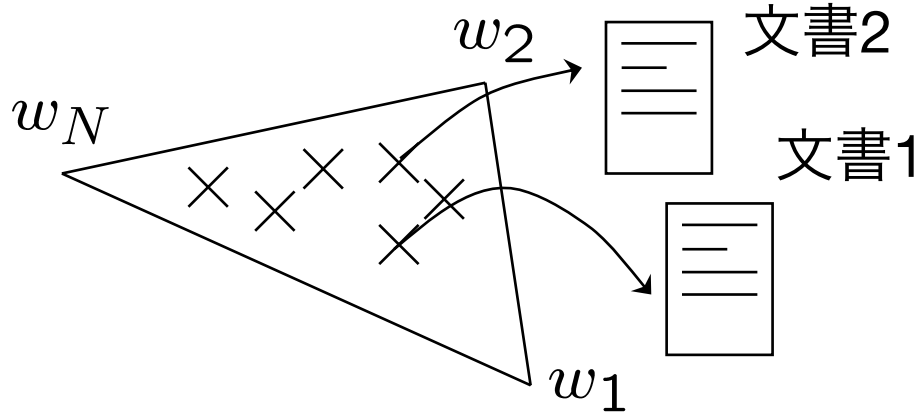
持橋大地
NTTコミュニケーション科学基礎研究所
daichi@cslab.kecl.ntt.co.jp

統計関連学会連合大会2009
「Bayes統計モデルのための計算技法とその応用」

自己紹介

- 研究分野：統計的自然言語処理
- 統計的自然言語処理とは
  - 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）
    - 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 …



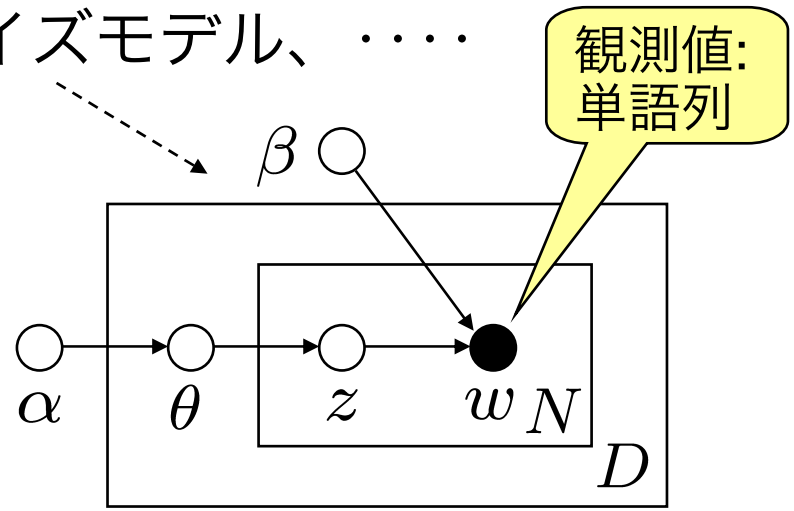


統計的自然言語処理

- 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



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|、|…

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

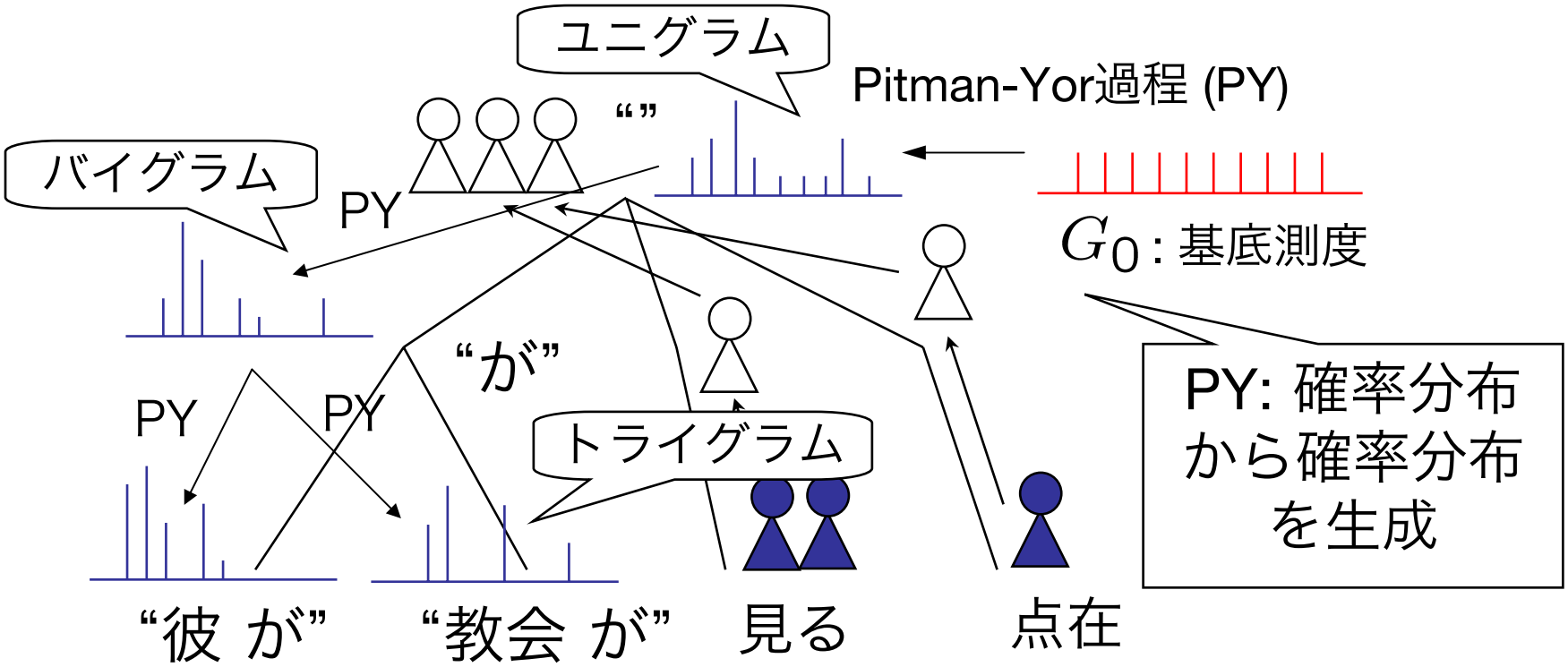
p(今日 は もう 見た)
= p(今日|^)・p(は|今日)・p(もう|は)・p(見た|もう)

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

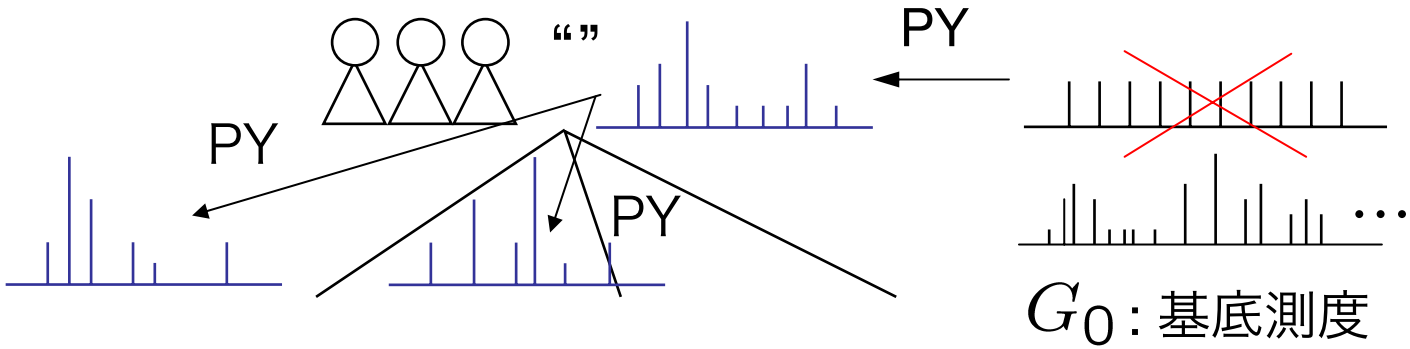
➡ 階層ベイズモデル: 階層Pitman-Yor過程言語モデル (HPYLM) (Teh 2006; Goldwater+ 2005)
- Pitman-Yor過程: ディリクレ過程 (GEM分布) の一般化

準備: HPYLM n-gram



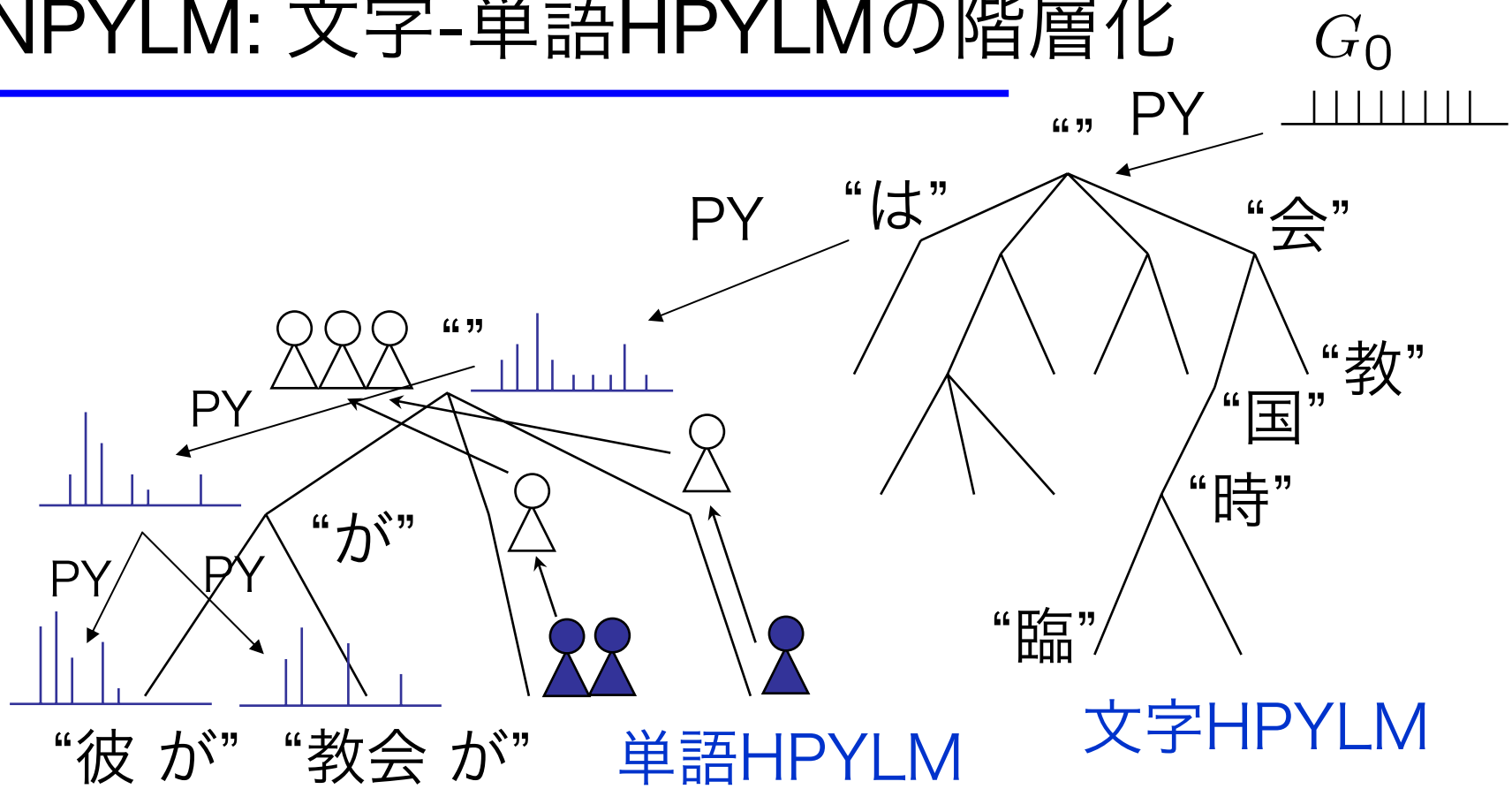
- カウントが0でも、より低いオーダーのMarkovモデルを用いて階層ベイズでスムージング
  - 注目している単語がそもそも存在しなかったら?

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

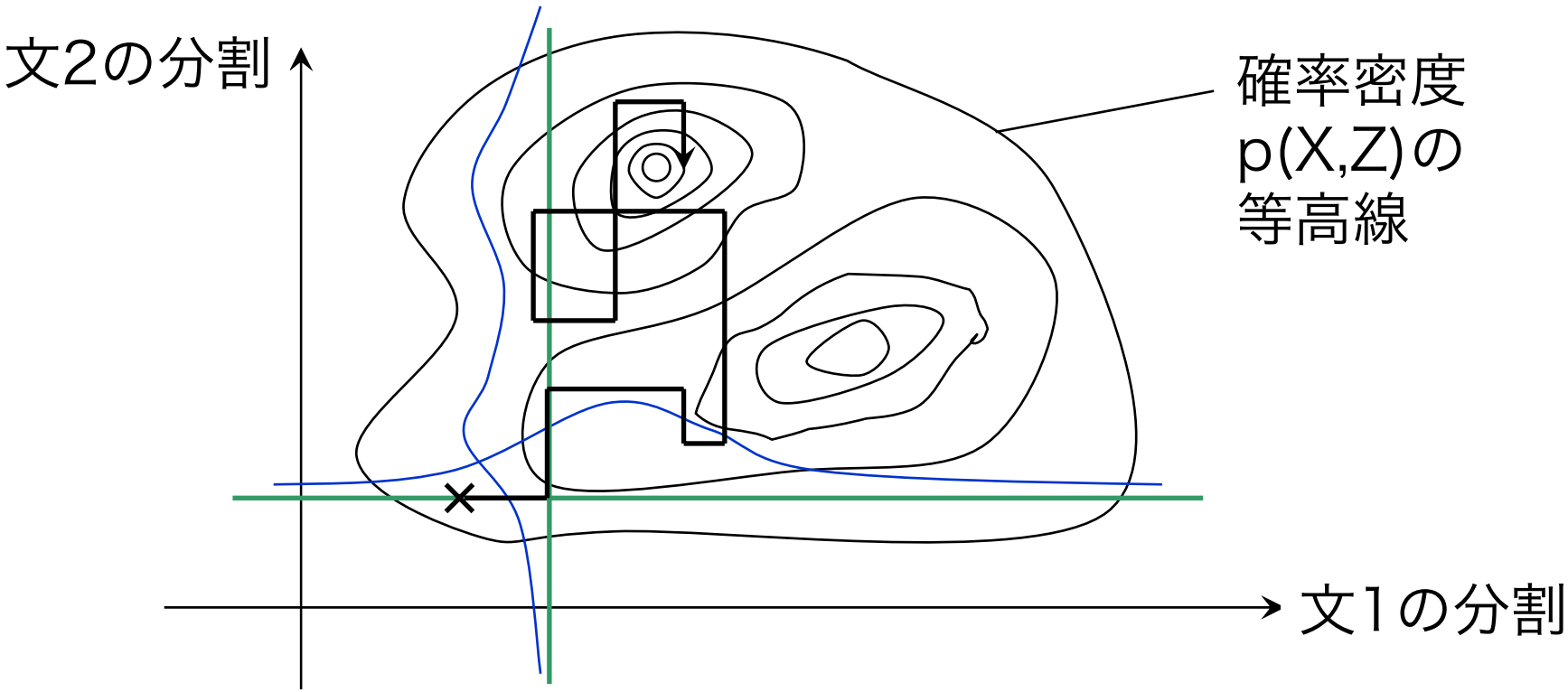
- データ: $\mathbf{X} = \{s_1, s_2, \cdots, s_X\}$ (文の集合)
  - 文: $s = c_1 c_2 \cdots c_N$ (文字列)
  - 隠れ変数: $\mathbf{z} = z_1 z_2 \cdots z_N$ ($z_i = 1$ のとき単語境界)
    - 隠れ変数の組み合わせは指数的に爆発
- 文がそれぞれ独立だと仮定すると、

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

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

  - 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

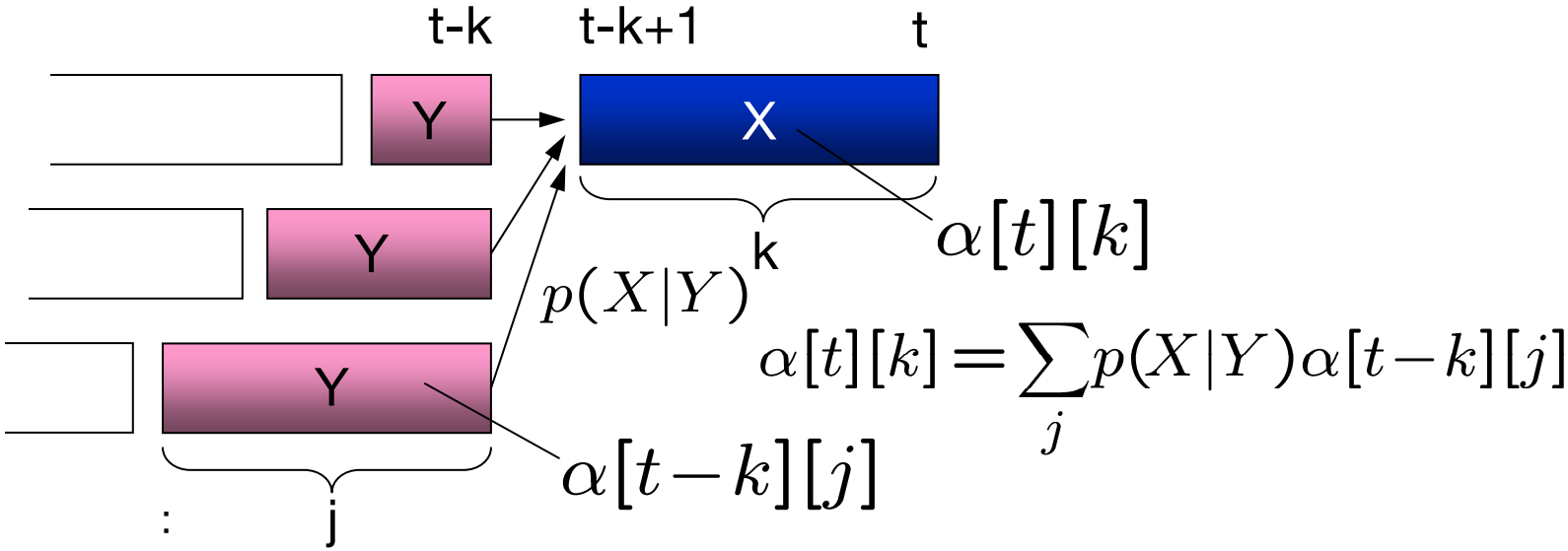
Gibbs Samplingと単語分割

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 。

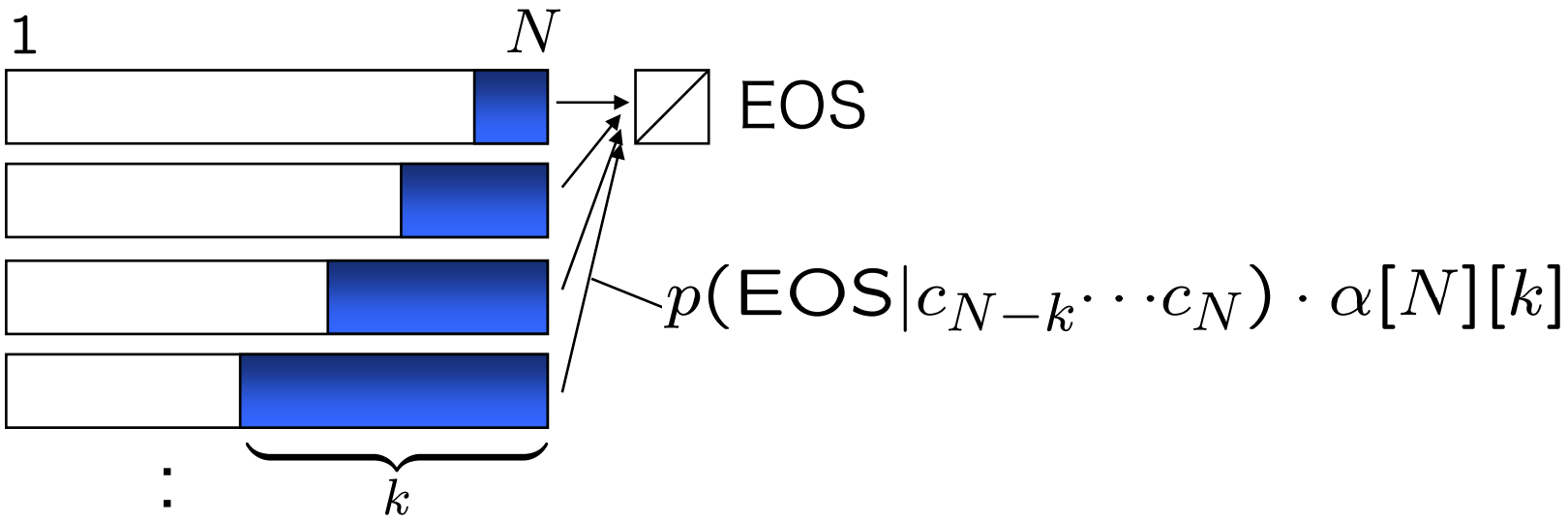
- ギブスサンプリングを繰り返すごとに、単語分割とそれに基づく言語モデルを交互に改善していく。

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

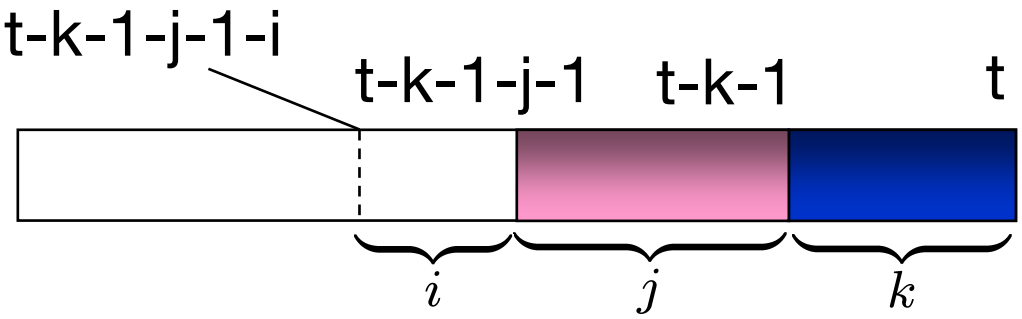


動的計画法によるデコード



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

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 。

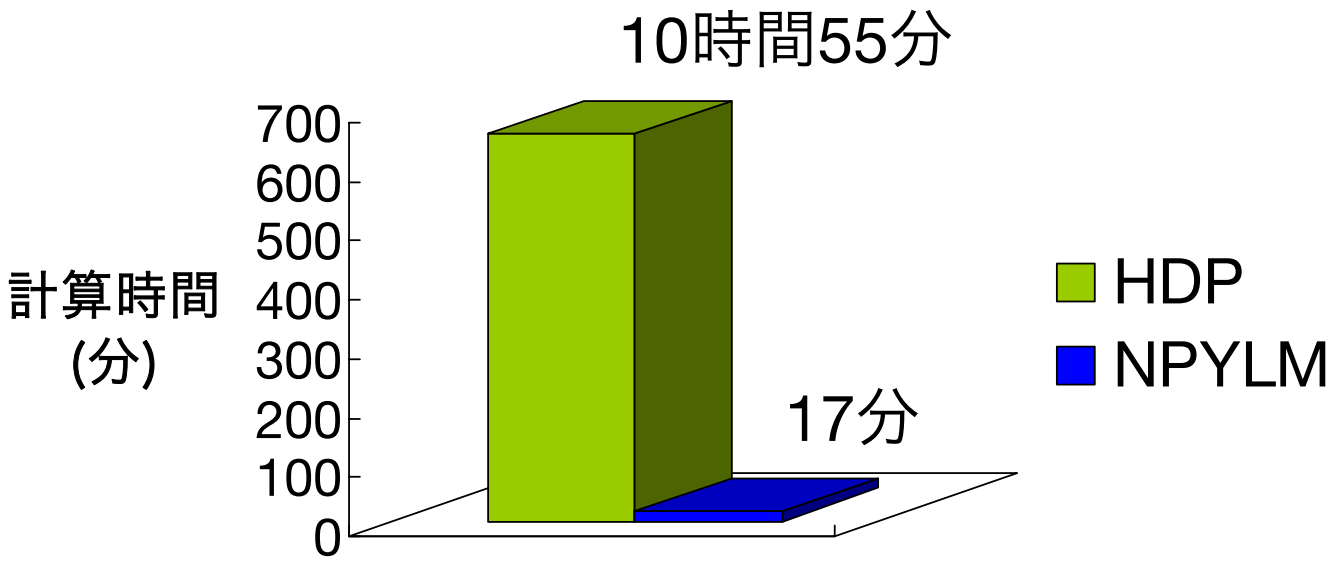
ビタビアルゴリズムで効率的に計算可能
(先行研究では不可能)

“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‥‥

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

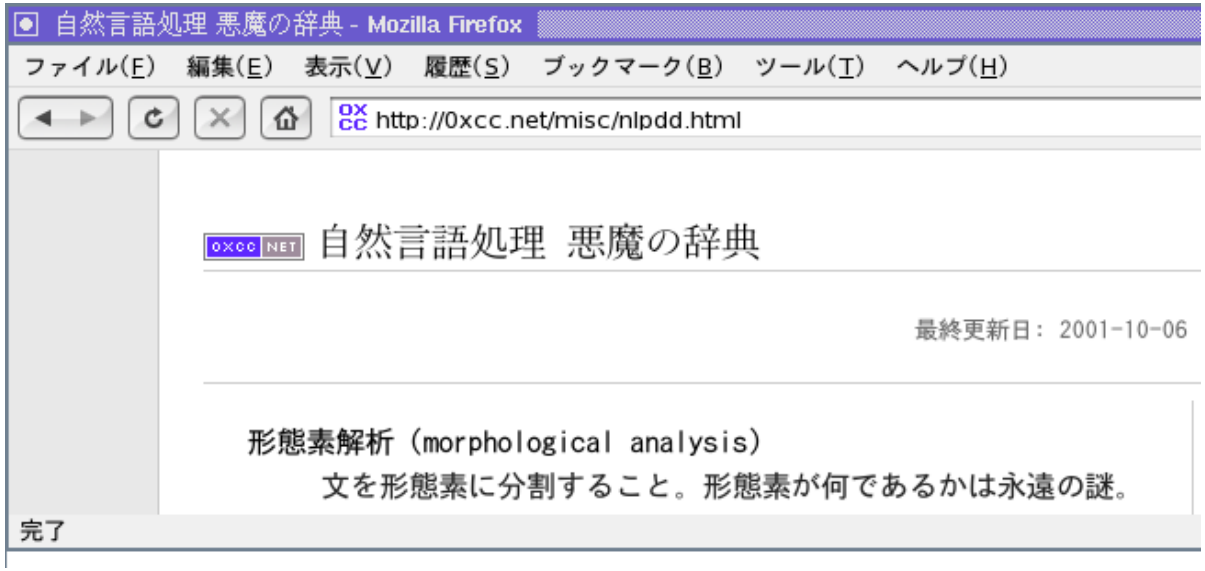
Google translate:
“Palestinian supporters of the event because of the Islamic Resistance Movement, Hamas.”

“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 …

“形態素”の再定義



- “自然言語処理 悪魔の辞典”: 高林哲氏
  - 「形態素が何であるかは永遠の謎」

教師あり学習では、確かに謎

- 今や謎ではない！
  - “形態素”とは、文字列の生成確率を最大にするような統計的な単位として導くことができる.

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

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