
ベイズ階層言語モデルによる 教師なし形態素解析

NTTコミュニケーション科学基礎研究所
持橋大地

daichi@cslab.kecl.ntt.co.jp

IPSJ SIGNL 190
2009-3-25 (水)

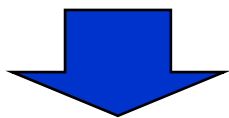
現在の形態素解析は完全か？

MeCabでの解析例

- ・ 前|スレ|1000|とりの|が|し|た|\(^|o|^)/
 - ・ ち|い|、|忠臣蔵|の|あら|すじ|おぼえ|た|!
 - ・ いづれ|の|御|時|に|か|、|女御|更衣|あ|また|さ|ぶら|ひ|た|ま|ひける|中|に|、|…
- 形態素解析の精度は「99%以上」と言われているが
 - ..
 - 基本的に、新聞記事のみでの評価
 - 掲示板やブログの実際の文では、とても99%は無理
 - 音声認識、話し言葉の「正しい」教師データ？
 - 未知の言語や古文には教師データが無い

教師なし形態素解析

- 次々に現れる新語、新表現をいちいち、人手で辞書登録するのか？
- 教師あり学習では、「単語分割の基準」がヒューリスティック
- 複雑な言語的知識はともかく、単語分割程度は教師なしで自動学習できるべきではないか？
 - 情報理論的な「単語」の基準を与えたい



教師なし形態素解析.

- ・ 教師データを使わない
- ・ 辞書を使わない
- ・ 自然言語一般の統計モデル

教師なし形態素解析: これまでの研究

- ヒューリスティックな基準
 - 連続する文字列の生起の検定が有意か
 - 前接/後続する文字分布のエントロピー (Jin, 田中2006)
 - MDLを用いた文字のチャンキング (松原 2007) etc..
 - .. 統計的意味が曖昧/一部の情報しか使っていない



- 確率モデルに基づく統計的定式化
 - 文字列 s を単語分割した確率


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

を最大化する、「言語として自然な」単語分割 \mathbf{w} を求める (永田 1996(教師あり); Goldwater+ 2006)

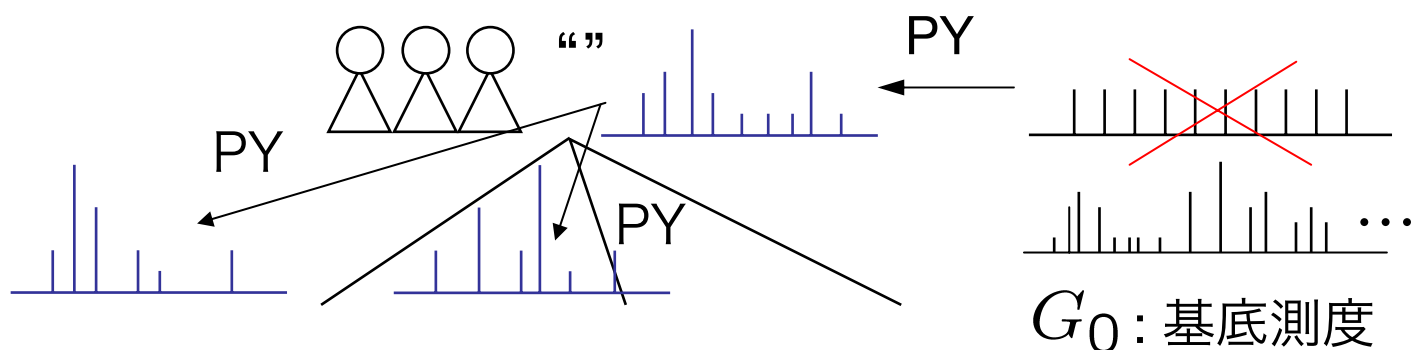
ヒューリスティックな基準も統計的に内包している

今回のアプローチ

$\operatorname{argmax}_{\mathbf{w}} p(\mathbf{w}|s)$: 言語モデル確率

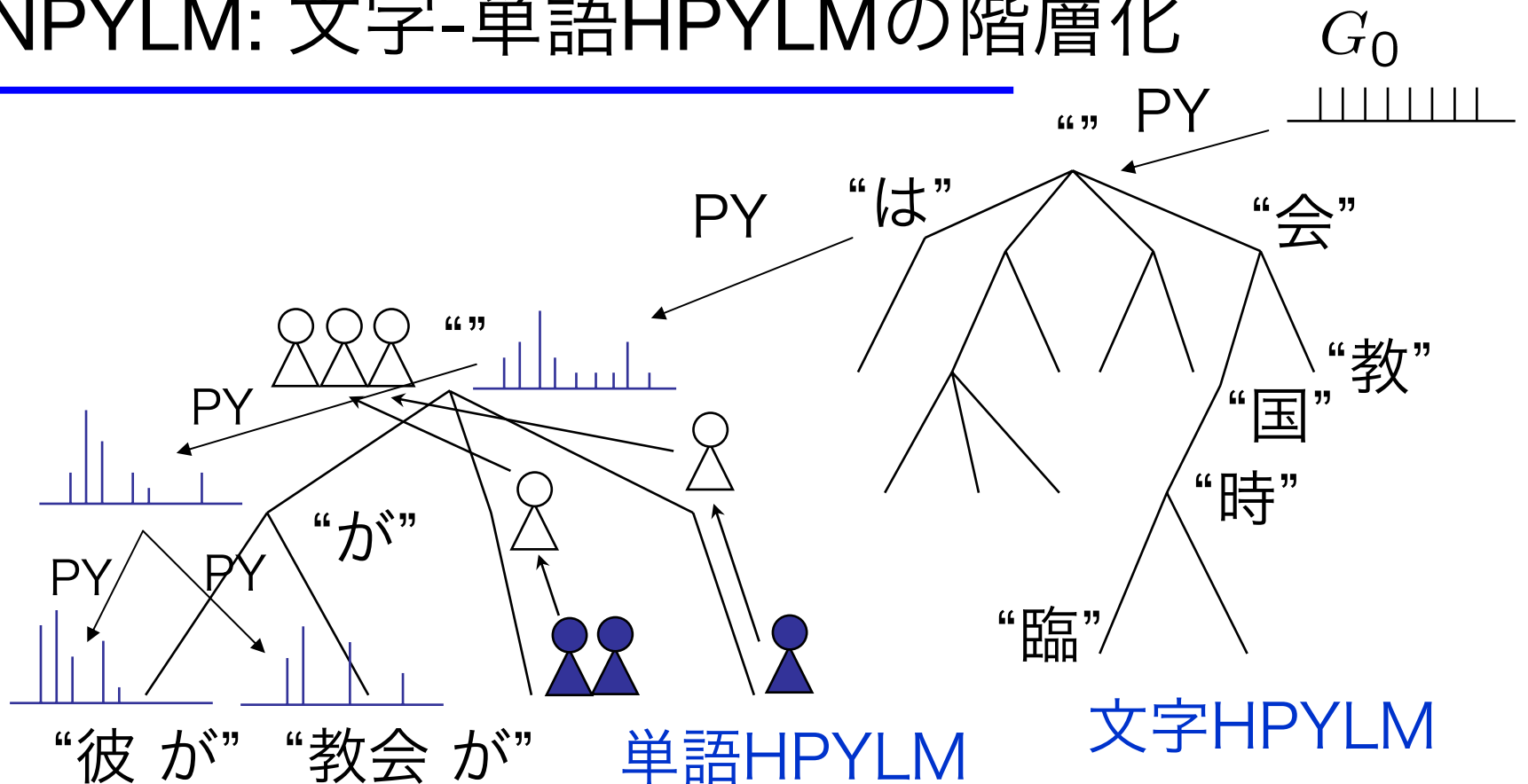
- 直接、ベイズ単語nグラム-文字nグラム言語モデルの性能を最適化する単語分割
 - 統計的意味が明確  結果的に直感とも一致
 - 文字あたりパープレキシティ最小化
- NPYLM: Nested Pitman-Yor Language Model
 - 文字列から、隠れた単語分割を推定しつつ直接言語モデルを作成できる
 - Byproductとして、形態素解析が可能
 - 未知の言語を含む、あらゆる言語に適用可能
 - HPYLM(ベイズn-gram言語モデル)の拡張

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
 - ME等で与えてもよい(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: $\mathbf{X} = \{s_1, s_2, \dots, s_X\}$ (文の集合)
 - 文: $s = c_1 c_2 \dots c_N$ (文字列)
 - 隠れ変数: $\mathbf{z} = z_1 z_2 \dots z_N$ ($z_i = 1$ のとき単語境界)
 - 隠れ変数の組み合わせは指数的に爆発

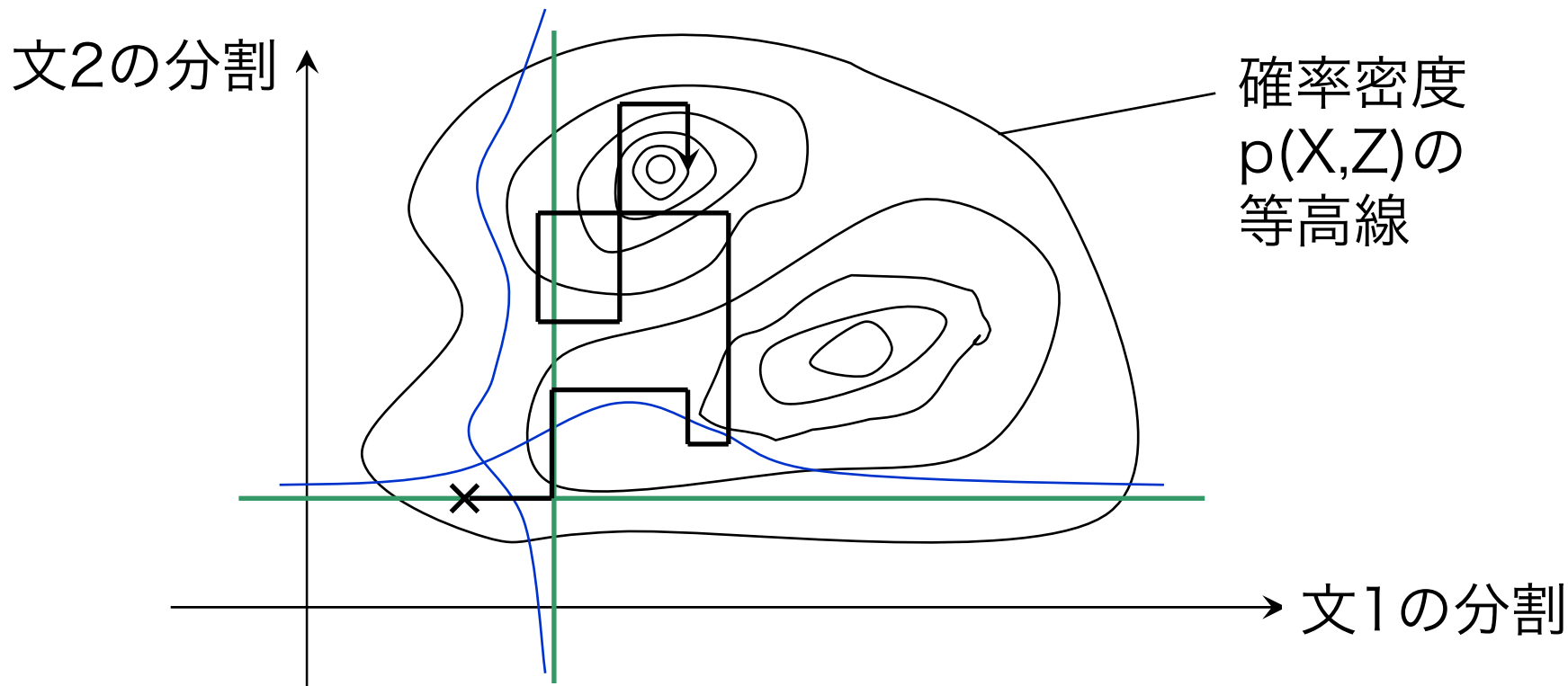
- 文がそれぞれ独立だと仮定すると、

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

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

- 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.

Gibbs Samplingとは



- データ X の確率を最大化する隠れ変数 Z を $p(X,Z)$ からサンプリングする方法
 - $p(z_i|X)$ からのサンプリングを充分繰り返すと、正しい解に収束

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 $s = s_1 \dots s_X$ do

$\text{parse_trivial}(s, \Theta)$.

← 文字列全体が一つの「単語」

1. For $j = 1 \dots M$ do

 For $s = \text{randperm}(s_1 \dots s_X)$ do

 言語モデルから $\text{words}(s)$ を削除

$\text{words}(s) \sim p(w|s, \Theta)$ をサンプリング

 言語モデルに $\text{words}(s)$ を追加して更新

done.

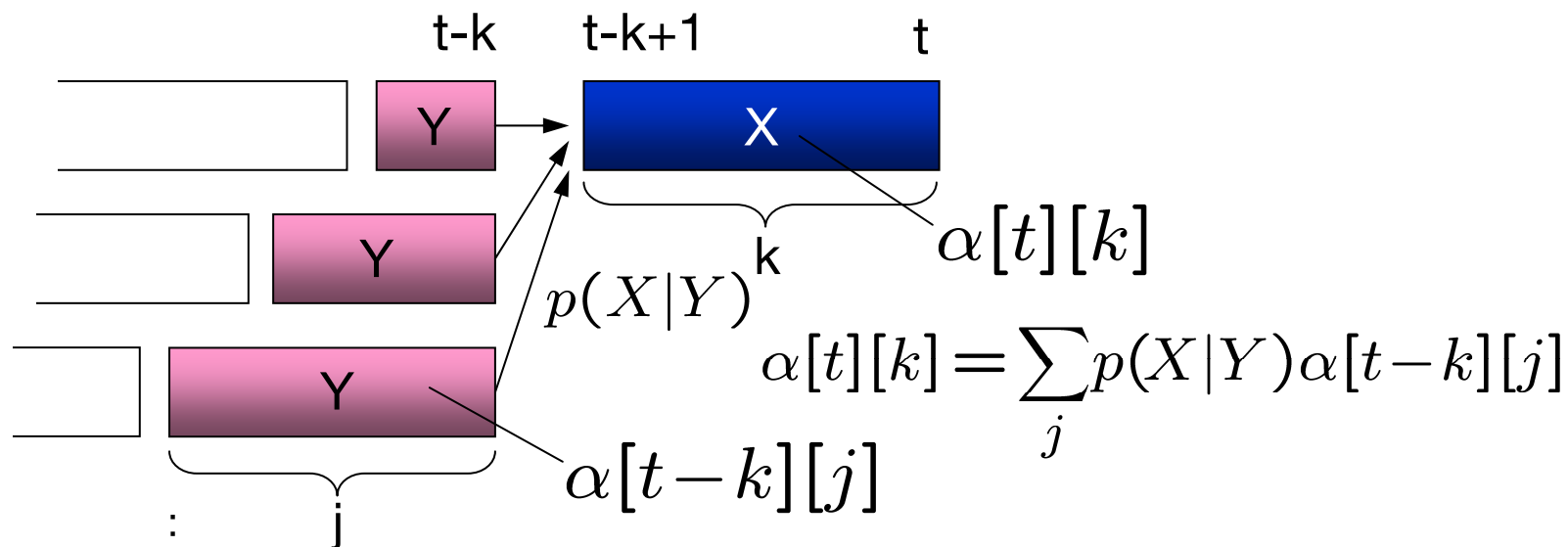
← Θ : 言語モデルのパラメータ

Gibbs Samplingと単語分割

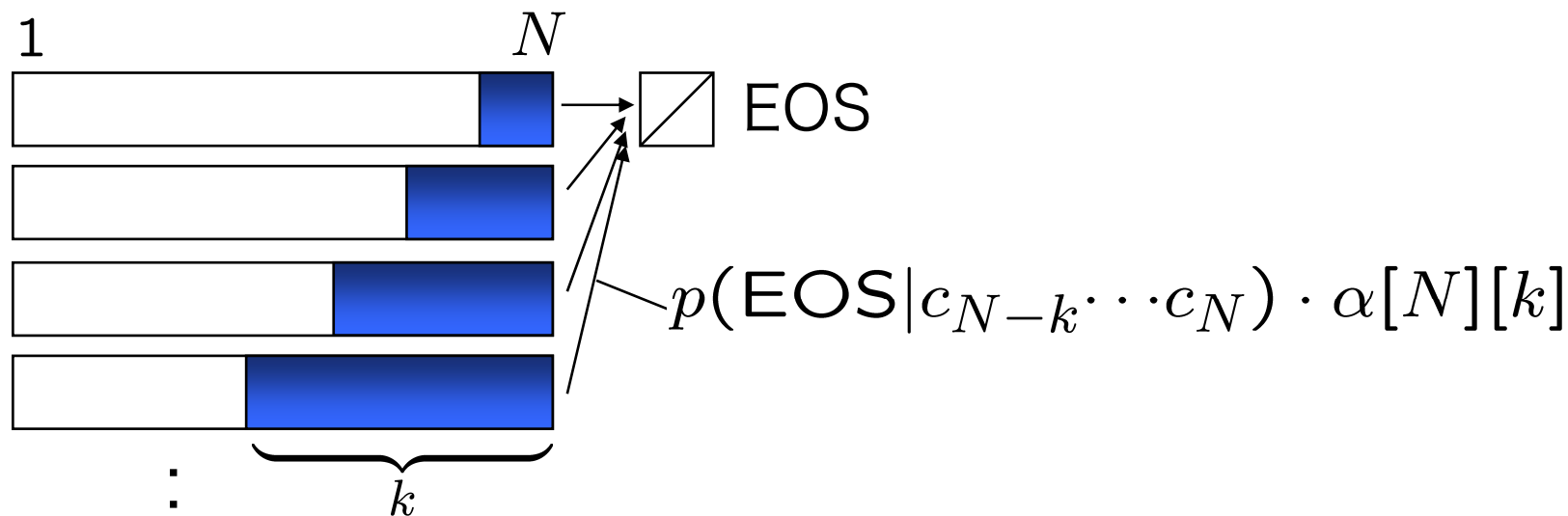
- 1 神戸では異人館 街の 二十棟 が破損した。
 - 2 神戸 では 異人館 街の 二十棟 が破損した。
 - 10 神戸 では 異人館 街の 二十棟 が破損した。
 - 50 神戸 では異人 館 街 の 二十棟 が破損した。
 - 100 神戸 では 異 人館 街 の 二十棟 が破損した。
 - 200 神戸 では 異人館 街 の 二十棟 が破損した。
- ギブスサンプリングを繰り返すごとに、単語分割とそれに基づく言語モデルを交互に改善していく。

動的計画法による推論

- $\text{words}(s) \sim p(w|s, \Theta)$: 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 $c_1 c_2 \dots c_t$ 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

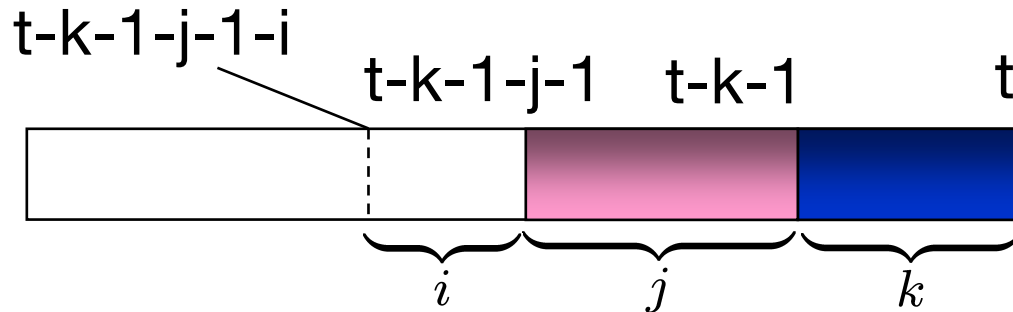


動的計画法によるデコード



- $\alpha[N][k]$ = 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
- $c_{N-k} \dots c_N$ 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、 $\alpha[t-k-1][j][i]$ ($i = 0 \dots L$) 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)

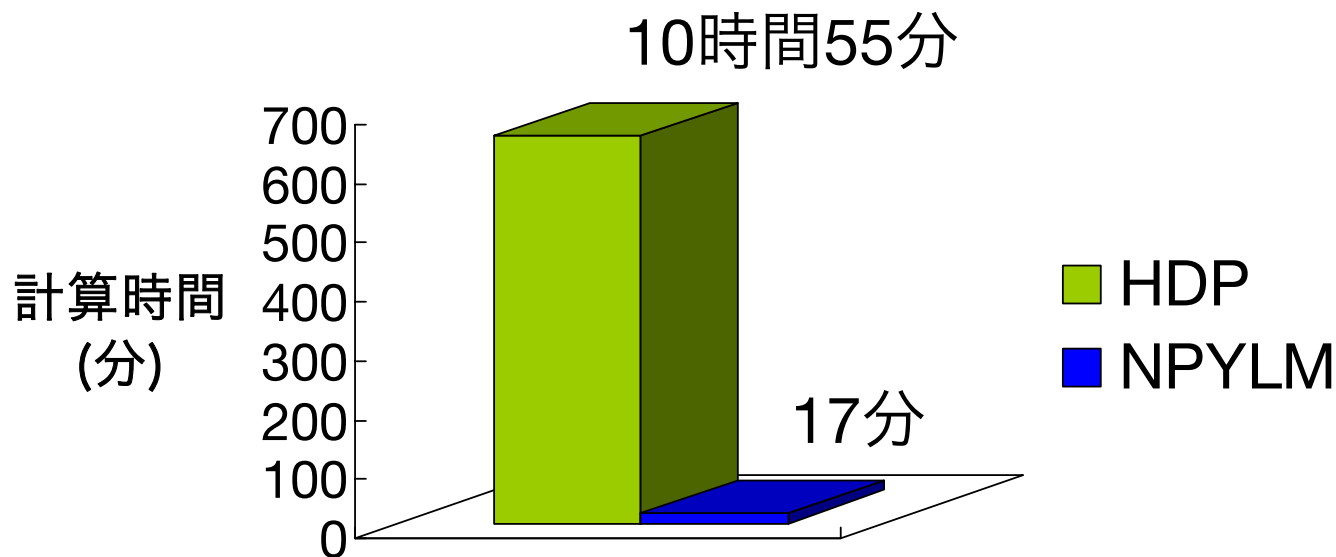
実験: 英語音素列データ

- Goldwater+(ACL 2006)のHDP単語バイグラムモデルとの比較
- 上で使われているCHILDES英語音素列データ
 - “WAtsDIs” → “WAts DIs” (What’s this) のように復元するタスク
- 結果: Precision, Recall, F値とも **非常に大きく改善**

モデル	P	R	F	LP	LR	LF
NPYLM	74.8	75.2	75.0	47.8	59.7	53.1
HDP	61.9	47.6	53.8	57.0	57.5	57.2

- 9,790文、平均9.8文字/文と少量のデータセット

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

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)
- 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

“正解”データとの違い

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。

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 — **齋田悟司 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**」 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。

「正解」データ

文法的判断

固有名詞を切りすぎない

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。

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 — **齋田 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**」 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。

「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)

الفلستيني بسبب تظاهرة لانصار حركة المقاومة الاسلامية حماس
و اذا تحقق ذلك فان كيسلو فسكيه قد حاز ثلاثه جري فيابرز ثلاثة

صحية
+قائد
الايقل

Google translate:

“Filstinebsbptazahrplansarhrkpalmquaompalaslami
iphamas.”

وقالت دانيل تومسون التي كتبت السيناريو. وقد استغرق اعداد خمسة اعوام. "تاريخي

↓ NPYLM

الفلستيني بسبب تظاهرة لانصار حركة المقاومة الاسلامية حماس
و اذا تحقق ذلك ف ان كيسلو فسكي يكون قد حاز ثلاثه جري فيابرز ثلاثة

صحية
سطينية
مالايقل

Google translate:

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

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

“Alice in Wonderland”の解析

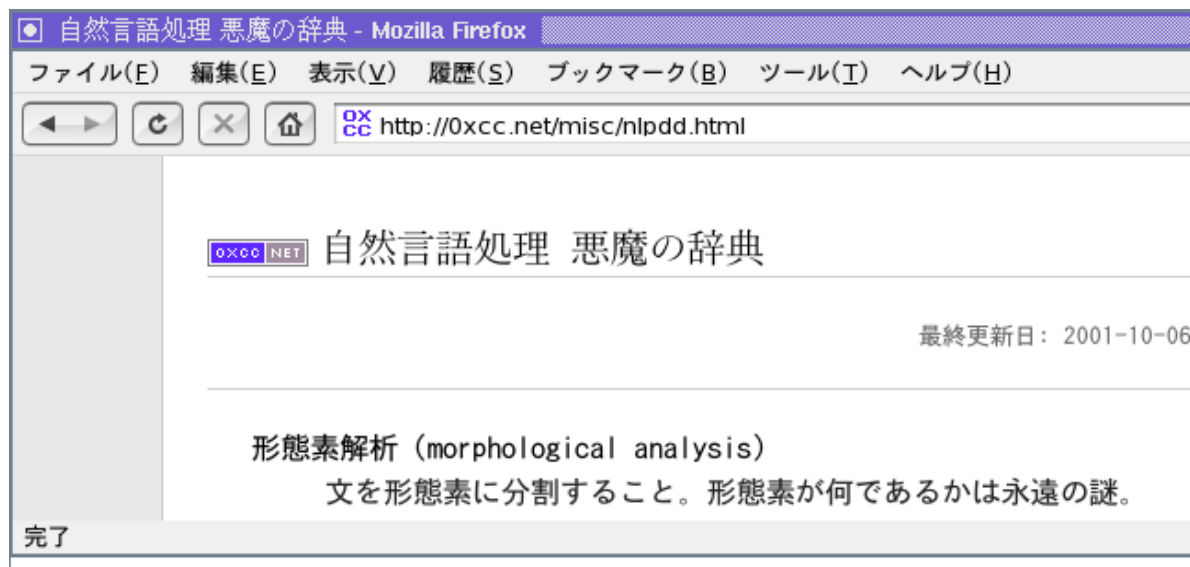


first, she dream ed of little alic e herself , and once again the tiny hand s were clasped up on her knee , and the bright eager eyes were looking up into hers -- she could hear the very tone s of her voice , and see that queer little toss of her head to keep back the wandering hair that would always get into hereyes -- and still as she listened , or seemed to listen , the whole place a round her became alive the strange creatures of her little sister 's dream. the long grass rustled at her feet as the white r abb it hurried by -- the frightened mouse splashed his way through the neighbour ing pool -- she could hear the rattle of the tea cups as the march hare and his friends shared their never -ending meal , and the shrill voice of the queen ...



first, she dream ed of little alic e herself , and once again the tiny hand s were clasped upon her knee , and the bright eager eyes were looking up into hers -- she could hear the very tone s of her voice , and see that queer little toss of her head to keep back the wandering hair that would always get into hereyes -- and still as she listened , or seemed to listen , the whole place a round her became alive the strange creatures of her little sister 's dream. the long grass rustled at her feet as the white r abb it hurried by -- the frightened mouse splashed his way through the neighbour ing pool -- she could hear the rattle of the tea cups as the march hare and his friends shared their never -ending meal , 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情報理論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Kneser-Ney nグラムを学習する方法ともみなせる

展望と課題

- 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

おわり

ご清聴ありがとうございました。