Prefix確率を用いたプラン認識のWebアクセスログ解析への応用

東京工業大学
情報理工学研究科　計算工学専攻
佐藤研究室 :小島 諒介

背景

Webサイトのサービスの多様化

http://www.amazon.com/

背景

閲覧者の目的の推定

- Webサイトの改善
- 目的に適した広告の表示

目的

Webサイト利用者のプラン認識

プラン認識：

エージェントの行動からエージェントの目的・プランを推定

考慮する点

- ユーザの行動は不確実

  → 確率モデルを利用

- 目的を推定し，広告表示に利用

  → オンラインで目的を推定

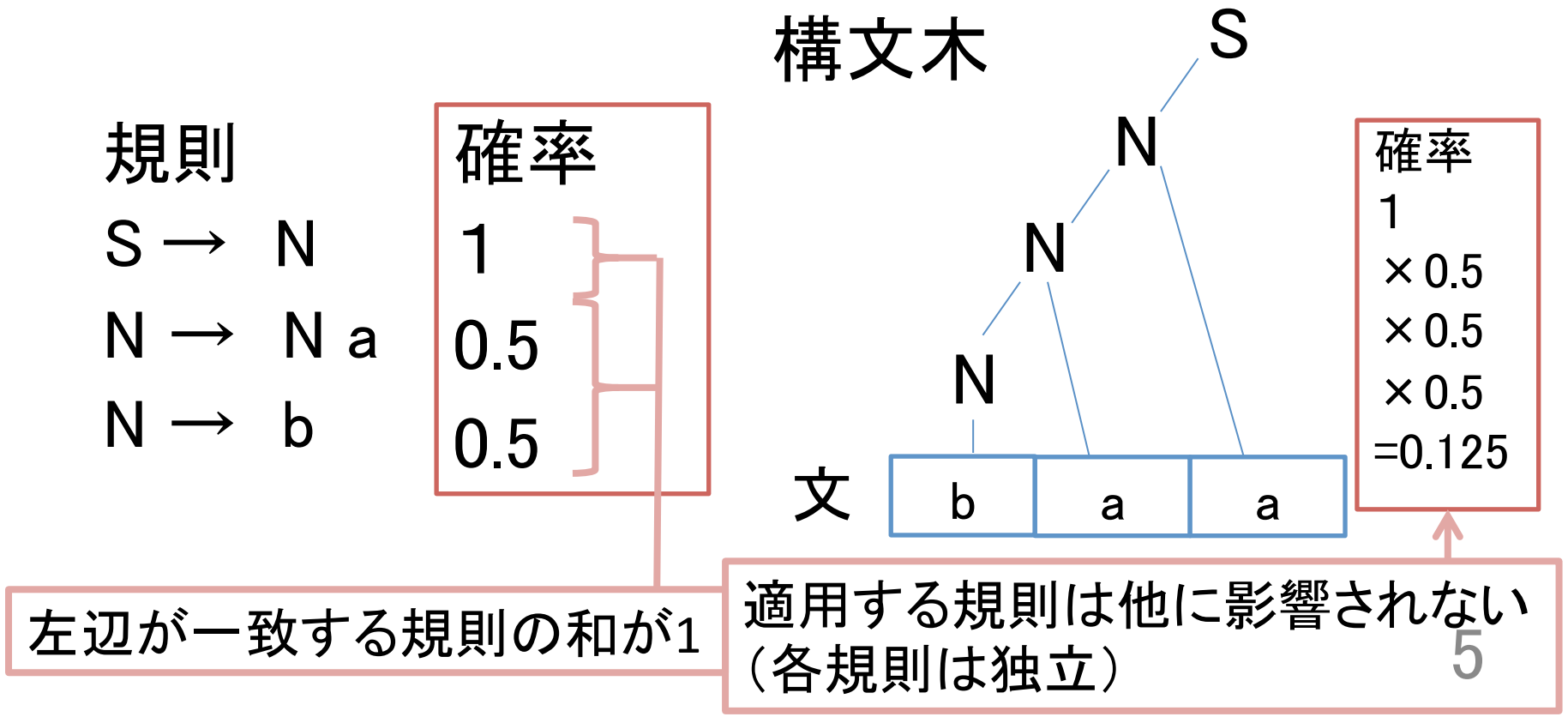
- Webサイトの改善のための分析に利用

  → 目的を達成していない（不満のある）ユーザの行動を扱う

確率文脈自由文法（ PCFG ）[Manning+99]

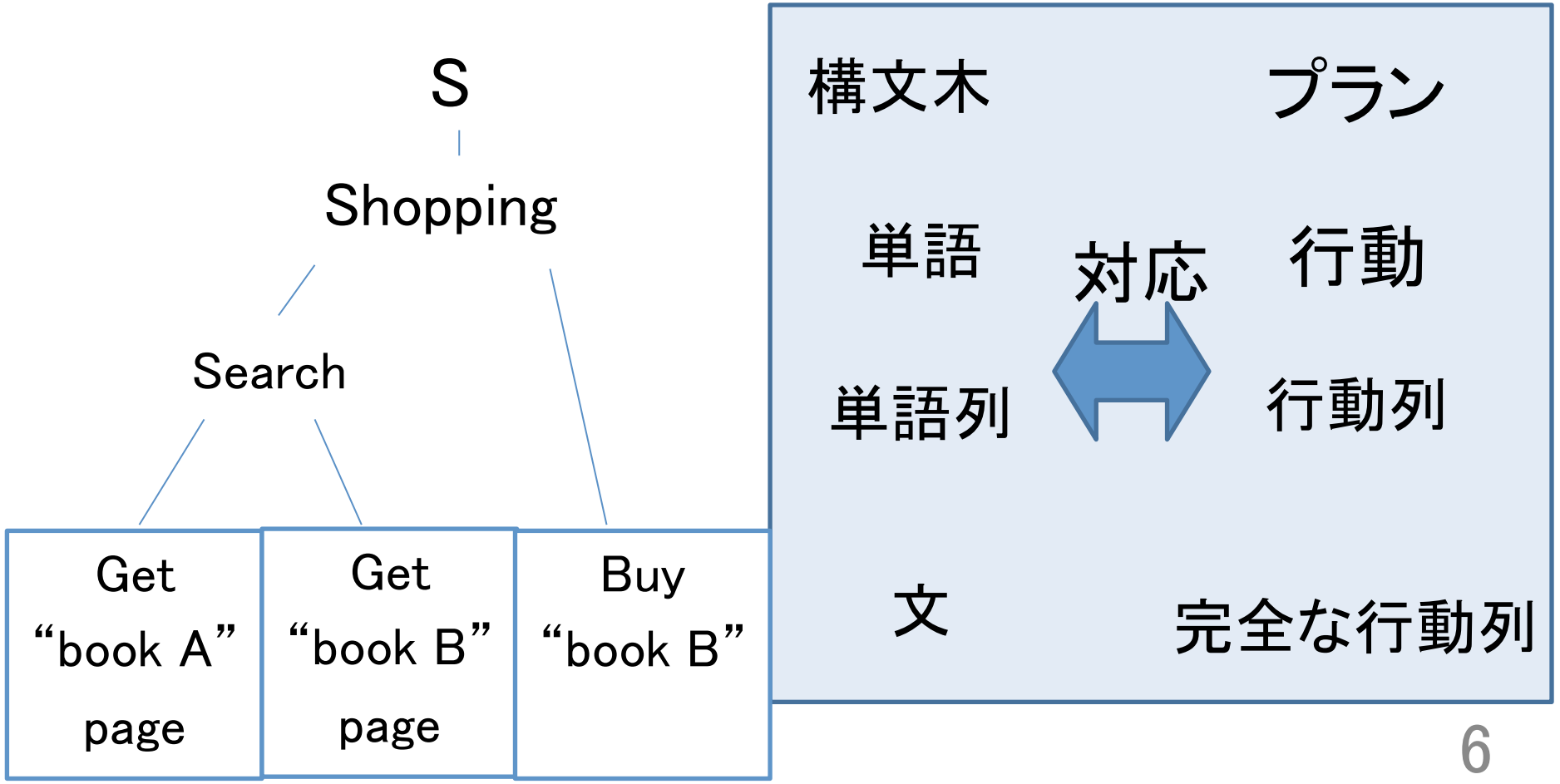
確率モデルの一つ

PCFGは各生成規則に確率が付与されている文脈自由文法CFG（contex-free grammar）



プラン認識と文法[Kautz+ 91, Vilain 90]

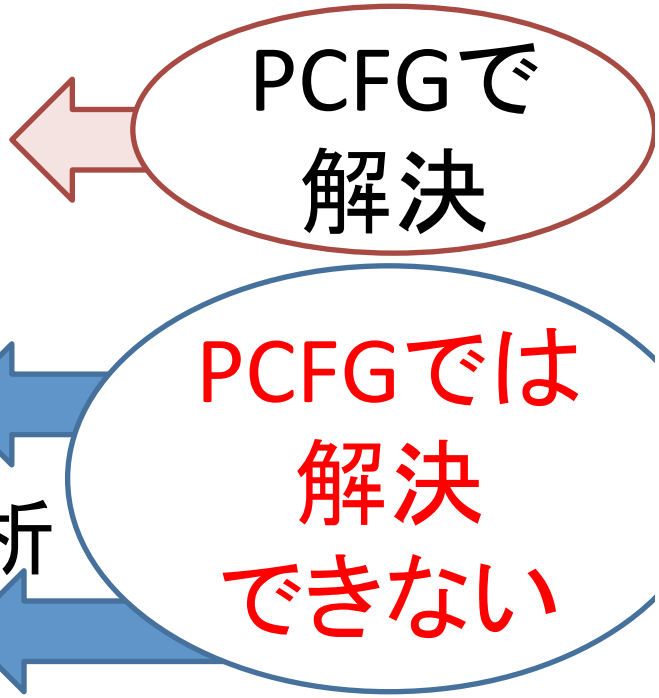
構文木をプランとして捉える



PCFGでは
文（完全な行動列）が与えられる必要がある

考慮する点

- ユーザの行動は不確実
  → 確率モデルを利用
- 広告表示
  → オンラインで目的を推定
- Webサイトの改善のための分析
  → 目的を達成していない
  ユーザの行動を扱う



提案法

完全な行動列の接頭部分列からプラン認識

○　prefix 確率 [Jelinek+ 91]を利用

Prefix: 文(完全な行動列)の接頭部分列

Prefix 確率:$Pr(x)$

同一 prefix xを持つすべての文の確率の和

例

有限の文 : abc, abcd, bcd

文の確率: $P(\text{“abc”}),\ P(\text{“abcd”}),\ P(\text{“bcd”})$

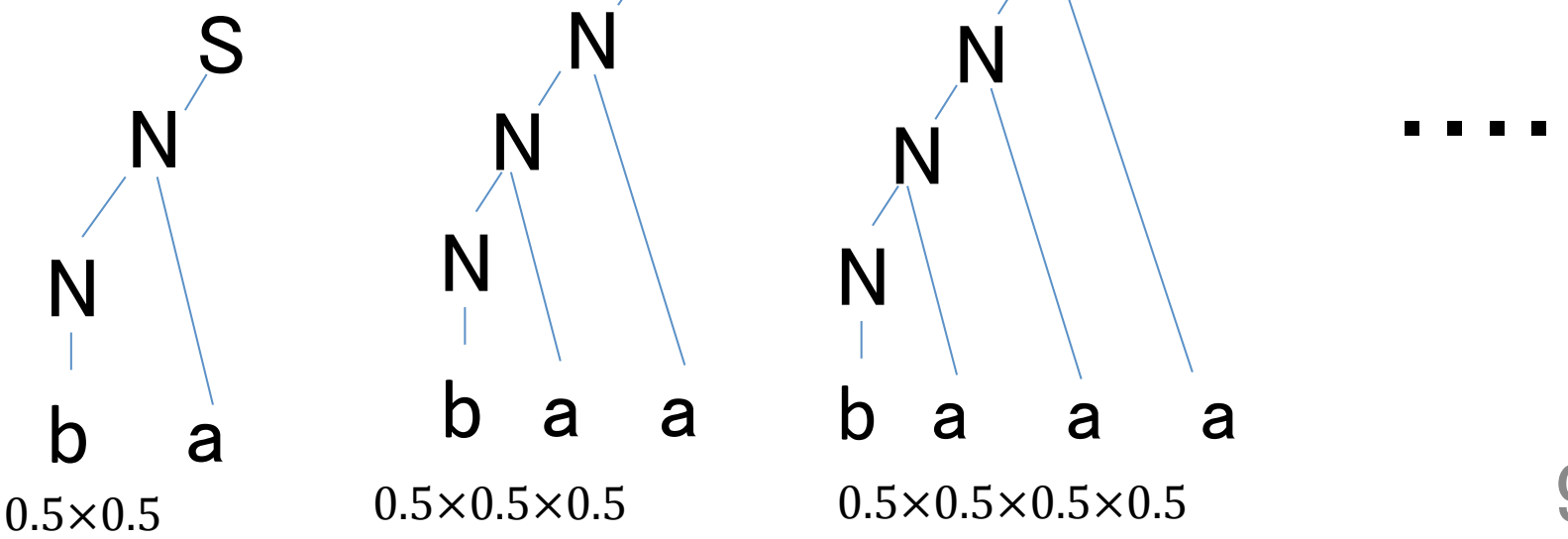
$Pr(\text{“ab”}) \equiv P(\text{“abc”}) + P(\text{“abcd”})$

prefix 確率の計算

同一 prefix を持つ文が無限に存在する場合がある

規則	確率
S → N	1
N → N a	0.5
N → b	0.5

prefix “b a” をもつ文



prefix 確率の計算

同一 prefix を持つ文が無限に存在する場合がある

規則	確率
S → N	1
N → N a	0.5
N → b	0.5

prefix “b a” をもつ文

$$Pr(\text{“}ba\text{”}) = (0.5)^2 + (0.5)^3 + (0.5)^4 + \ldots$$

$$= 0.5$$

提案法の利用

- 目的に応じた広告の表示
  - ユーザの(主たる)目的を推定し，利用する
- Webサイトの改善
  - ユーザのプラン全体を推定，利用する
  - →尤もらしい構文木の推定

提案法ではこれらを
Prefixから行う

疑問!

Prefix(完全な行動列の一部)からの推定結果は
(仮に)完全な行動列からの推定結果と一致するか？

評価実験：前準備

1. ユーザの行動
2. ユーザの（主たる）目的

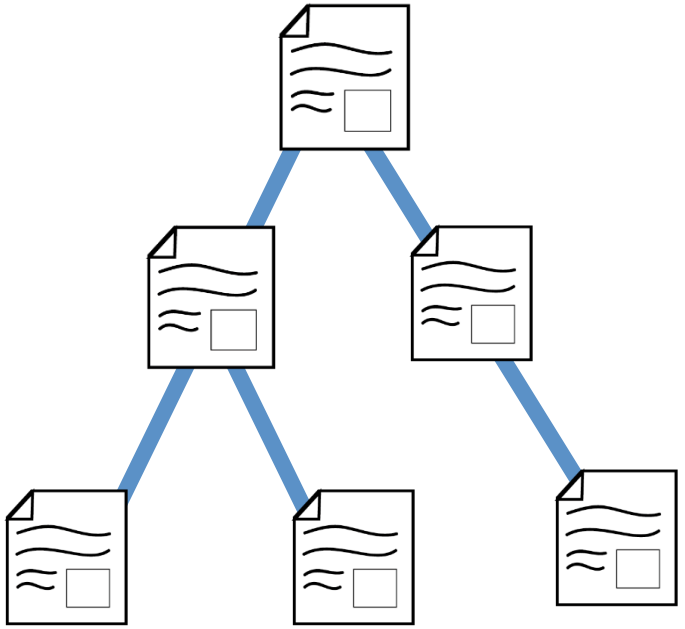
ユーザの行動

Webサイトの階層構造（ディレクトリ構造）上での行動を考える

up　　：階層構造を登る
down　　：下る
sibling：同じ階層の別ページへ移動
same　：同じページを再度要求する
move　　：その他の移動

行動列はこれらの記号の列

Webサイト



Webページのパスの例

/en/publication/index.html

ユーザの目的（予備実験のクラスタ分析の結果）

目的	主な行動
幅広い調査	ディレクトリ構造の上下に行動
ニュースの調査	ディレクトリ構造の上下に行動 +同じページへのアクセス
特定分野の調査	同じ階層のページへのアクセス
特定分野のニュースの調査	同じページへのアクセス
その他	ディレクトリ構造に沿わない行動

実験で使用する規則の例

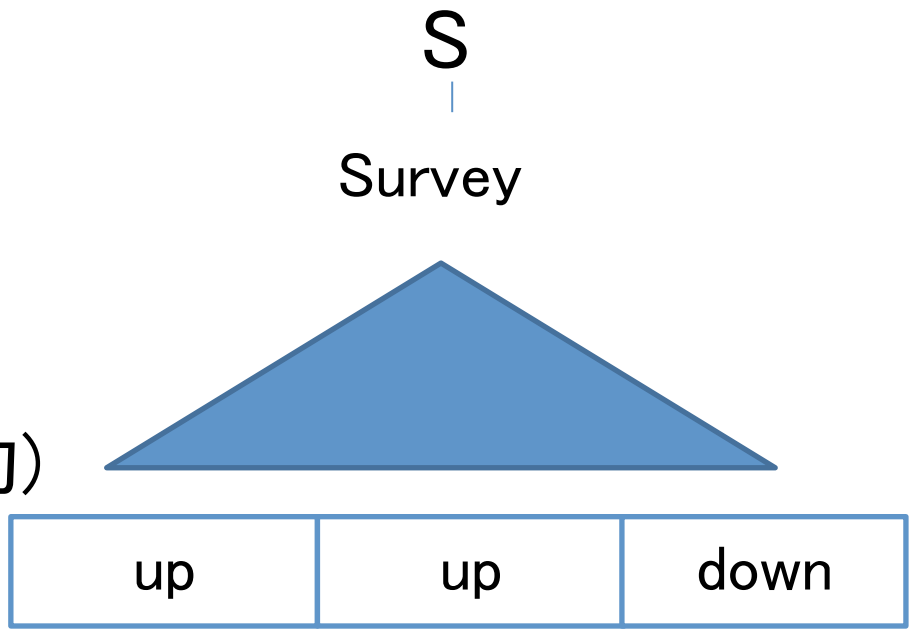
- ユーザの目的

  S → Survey

  S → News

  …

  （前で説明した5種類の目的）



- 繰り返し

  Up → Up, up

- 複数の動き

  UpDown →Up,Down

  UpDown →Up, SameLayer, Down

実験

- 実験の目的

  prefixから推定した目的

  完全な行動列から推定した目的

  } 一致するか？

- 実験データ:
  - 実験1. 人工アクセスログデータ
  - 実験2. 実アクセスログデータ

- 前処理
  - アクセスの失敗を除去
  - リソースへのアクセスを除去
  - 長さ20～30に制限

アクセスログの例

example.com
2013/07/26 11:00
sato-www.cs.titech.ac.jp/en/publication/

- 処理系:記号的確率モデリングシステム PRISM

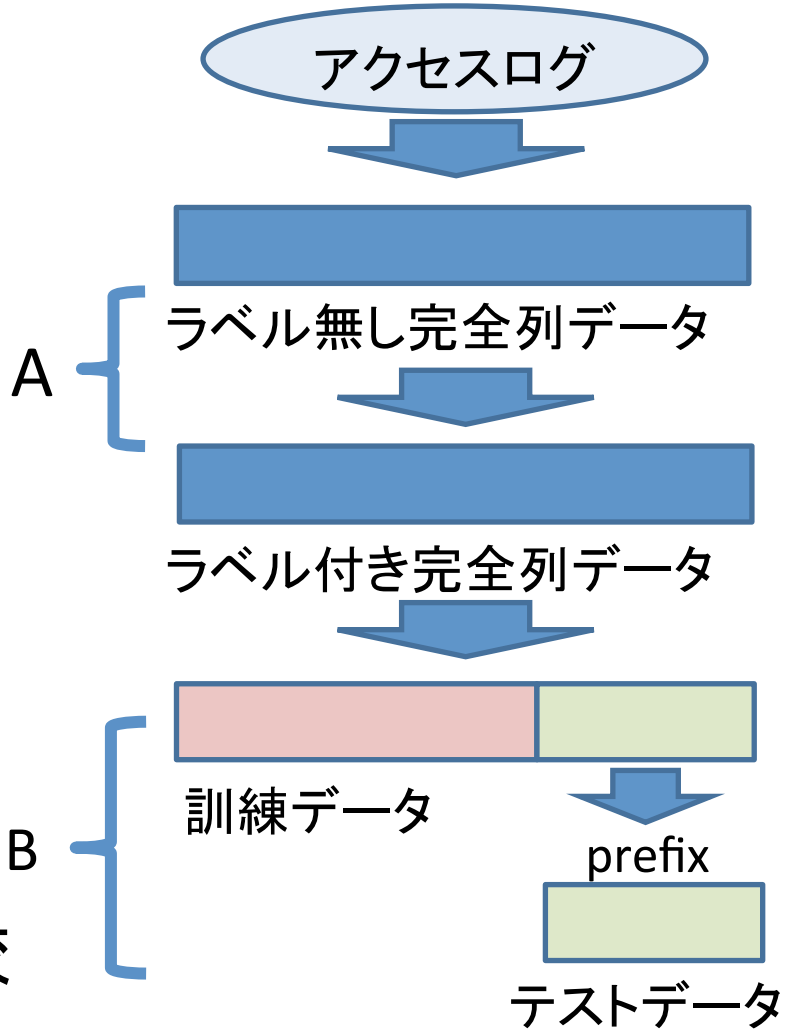
実験内容：概要

A. ラベル付与

- PCFGのパラメータ学習
- 完全文から推定した目的（正解ラベル）を付与

B. タスク,評価方法

- Prefix から目的を推定
- 予測ラベルと正解ラベルを比較（5-fold cross-validation）



実験1．人工アクセスログデータ

① データの性質を変えて実験
- HMMから生成
- PCFGから生成（簡易文法を利用）

データ数1000

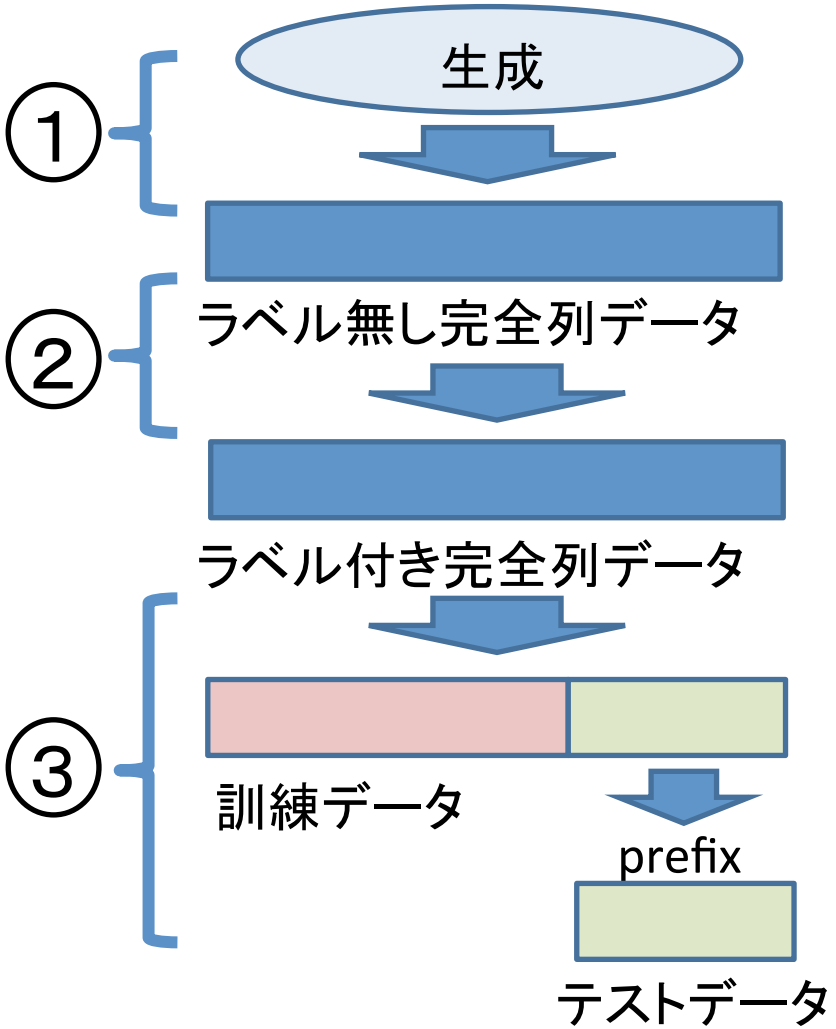
② （完全列）ラベル推定(PCFG)
- 複雑な文法

（ 規則：102個/ 非終端記号32個）
- 簡単な文法

（ 規則：42個/ 非終端記号24個）

③ （prefix）ラベル推定手法
- 提案法
- HMM（状態数2～8）

5-fold cross validation



実験1．人工アクセスログデータ

① データの性質を変えて実験
- HMMから生成
- PCFGから生成（簡易文法を利用）

データ数1000

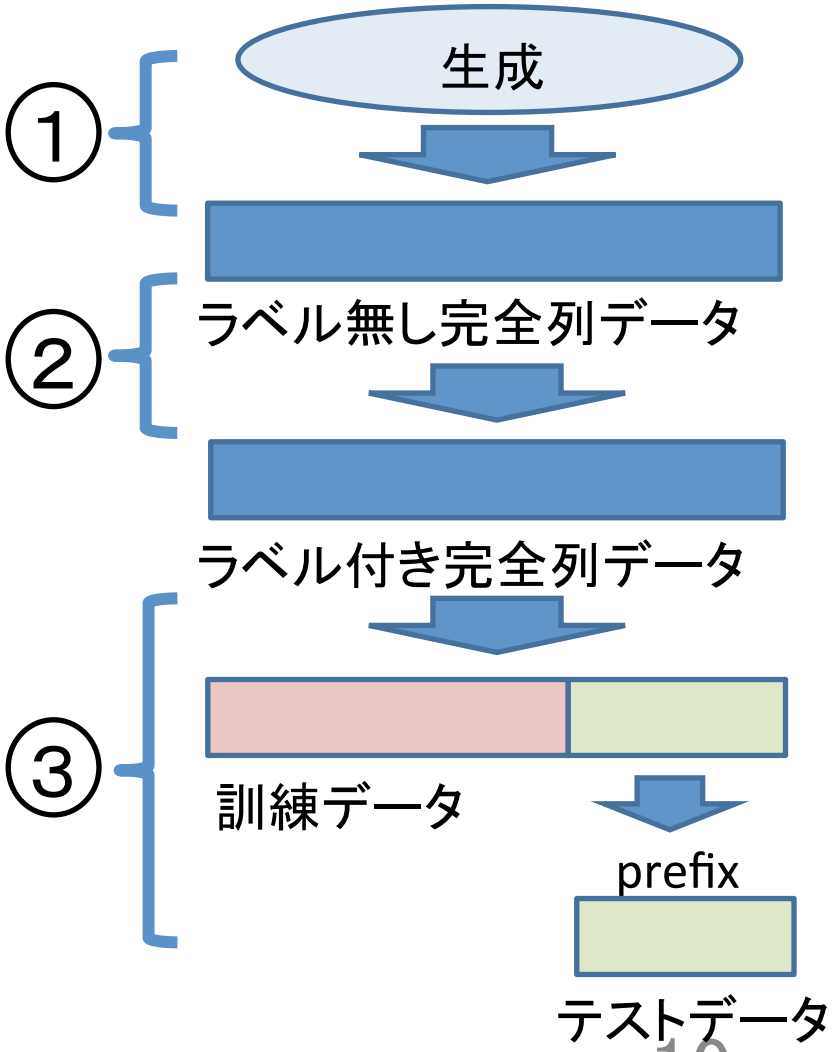
② （完全列）ラベル推定(PCFG)
- 複雑な文法

（ 規則：102個/ 非終端記号32個）
- 簡単な文法

（ 規則：42個/ 非終端記号24個）

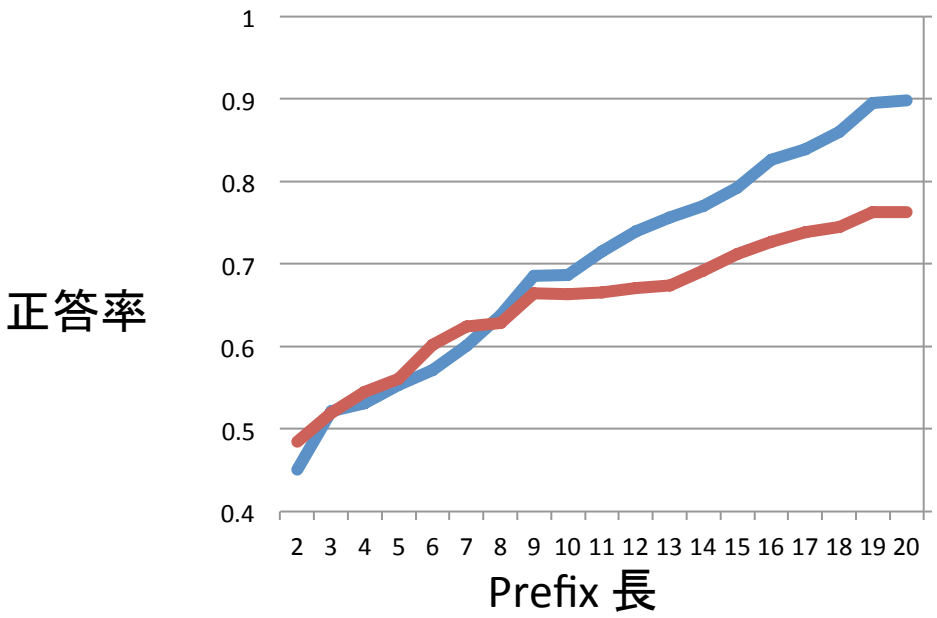
③ （prefix）ラベル推定手法
- 提案法
- HMM（状態数2～8）

5-fold cross validation

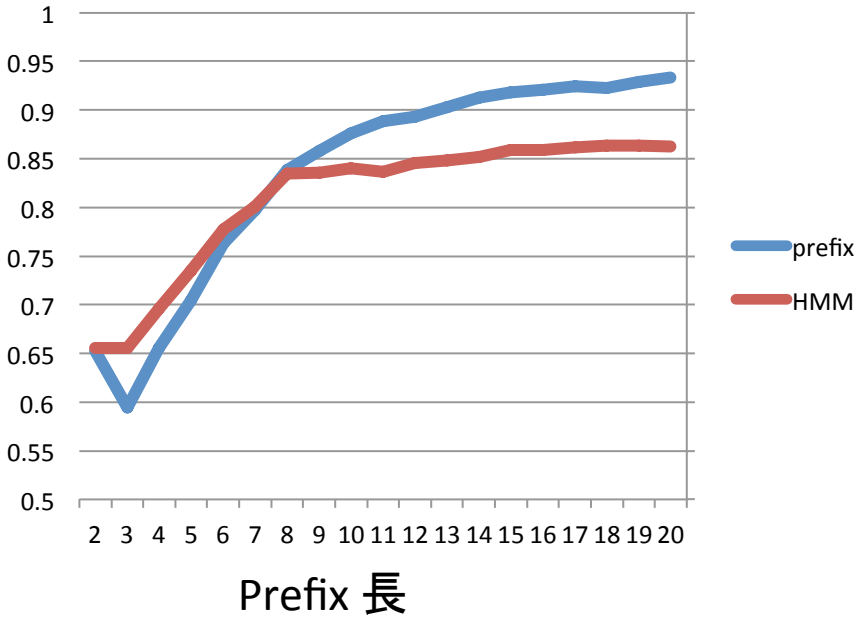


結果(1/2)

HMMから生成したデータに対する目的予測



PCFGから生成したデータに対する目的予測



- Prefix 長が長いとき提案法が有利
- HMMによる予測はPCFGから生成した規則性のある長い列に対して不利

実験1．人工アクセスログデータ

① データの性質を変えて実験
- HMMから生成
- PCFGから生成（簡易文法を利用）

データ数1000

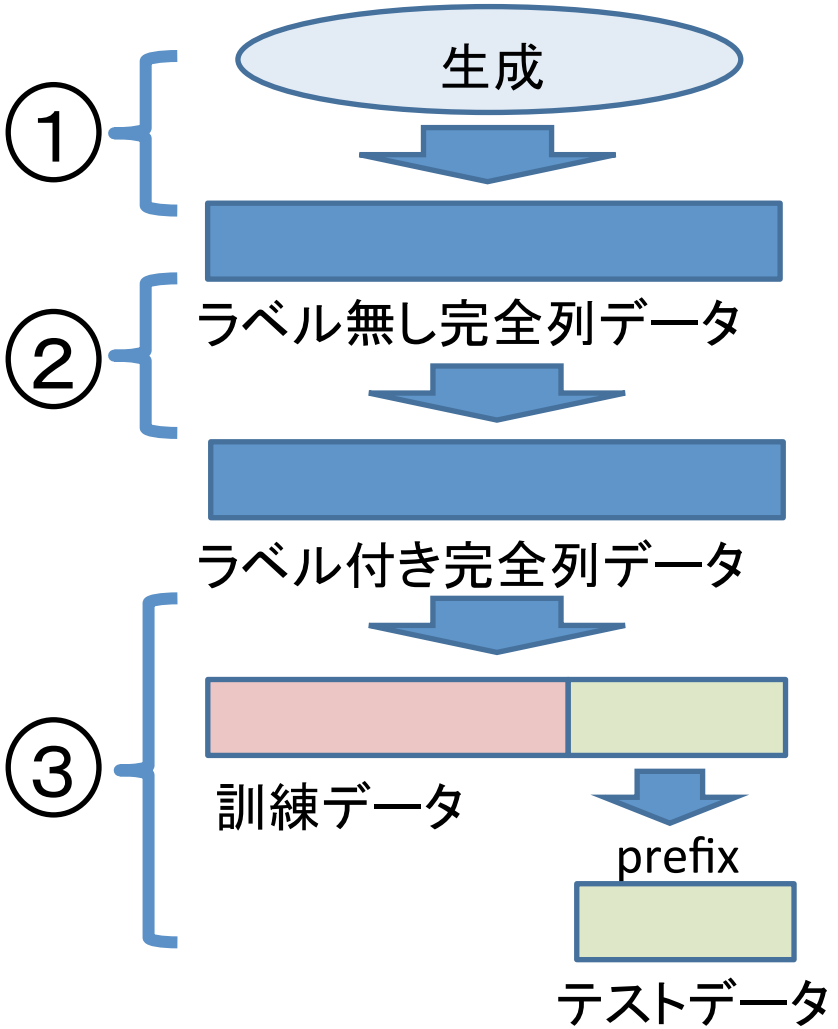
② （完全列）ラベル推定(PCFG)
- 複雑な文法

（ 規則：102個/ 非終端記号32個）
- 簡単な文法

（ 規則：42個/ 非終端記号24個）

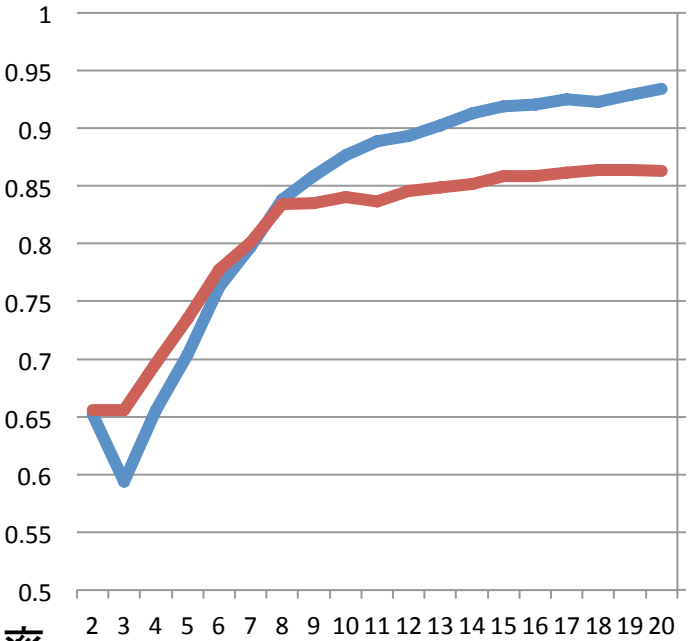
③ （prefix）ラベル推定手法
- 提案法
- HMM（状態数2～8）

5-fold cross validation

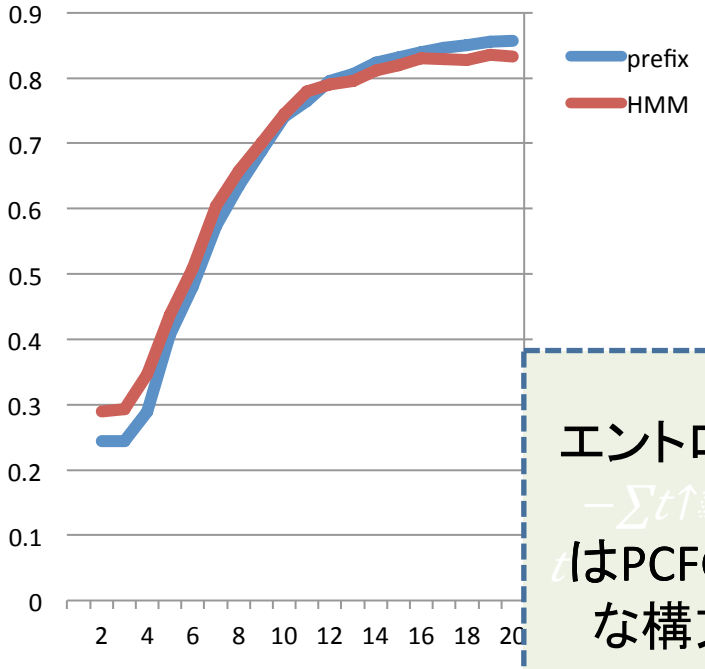


結果(2/2)

ラベルの付与に
複雑な文法を使用



ラベルの付与に
簡単な文法を使用



正答率

Prefix 長

エントロピーの定義
$-\sum t\uparrow\blacksquare p(t)\log p(t)$
はPCFGの導出可能な構文木[Chi99]

ラベルの付与に用いたPCFGのエントロピー

3.96×10↑6

2.09×10↑6

実験2. 実データ

the Internet Traffic Archive より3種類のデータを使用

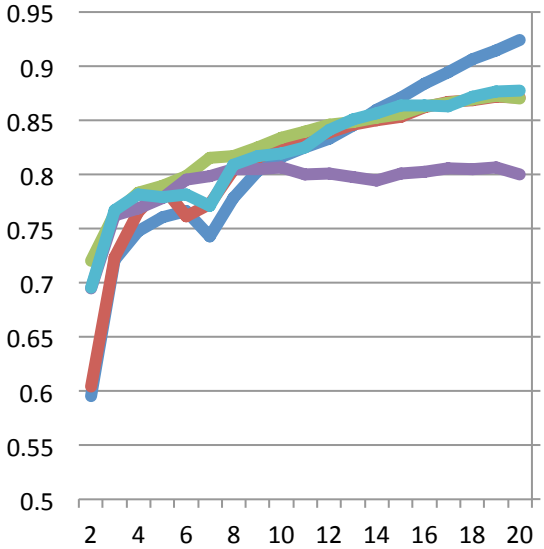
- U of S アクセスログ　　データ数： 652本
  (University of Saskatchewan)
- ClarkNet アクセスログ　データ数：4523本
  (An internet service provider)
- NASA アクセスログ　　データ数：2014本
  (NASA Kennedy Space Center)

HMMに加えその他の比較手法

ロジスティック回帰
SVM（行動列ベクトルをそのまま入力）
SVM （bag-of-words:行動の頻度）

結果

U of S

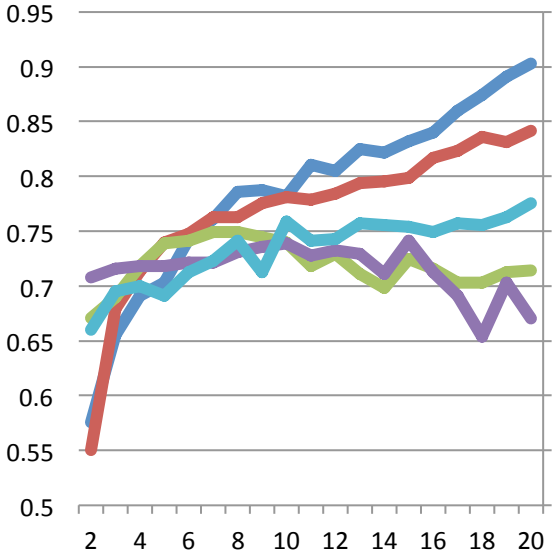


正答率

エントロピー

5.12×10^4

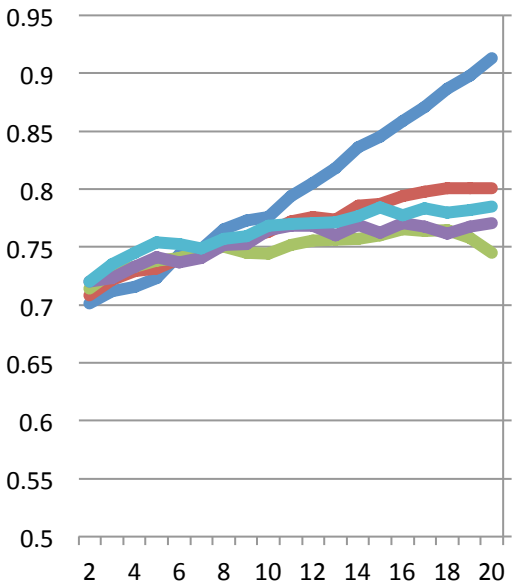
ClarkNet



Prefix 長

2.77×10^5

NASA



3.14×10^6

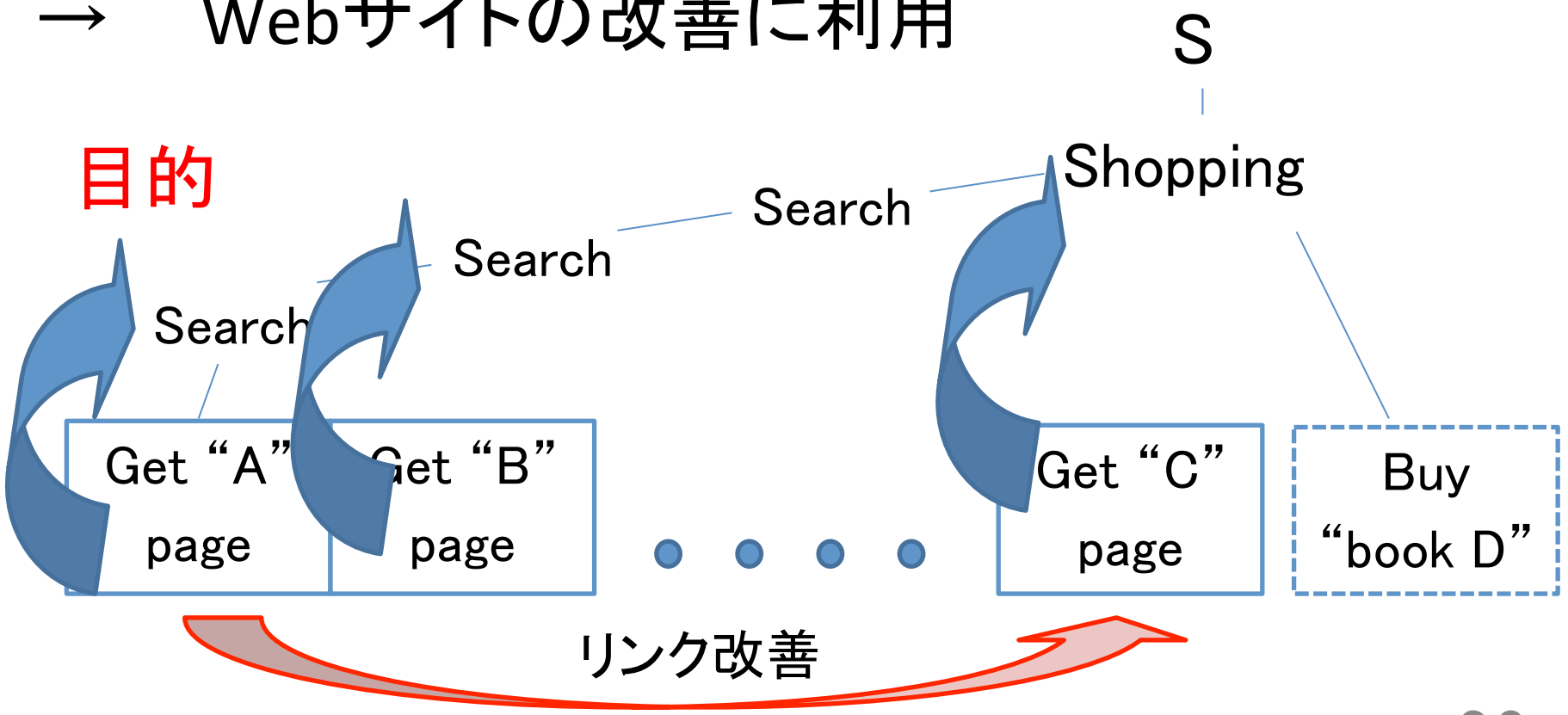
利用した文法はすべて同じ
(パラメータのみ異なる)

結論

- prefixが長いとき提案法の性能が良い
- 提案法が他の手法に比べ優位に働く場合の基準としてエントロピーが利用できる可能性がある
- prefixが短いとき提案法よりその他の手法の方が良い場合がある
  - ただし，最尤構文木の出力は提案法でしか利用できない

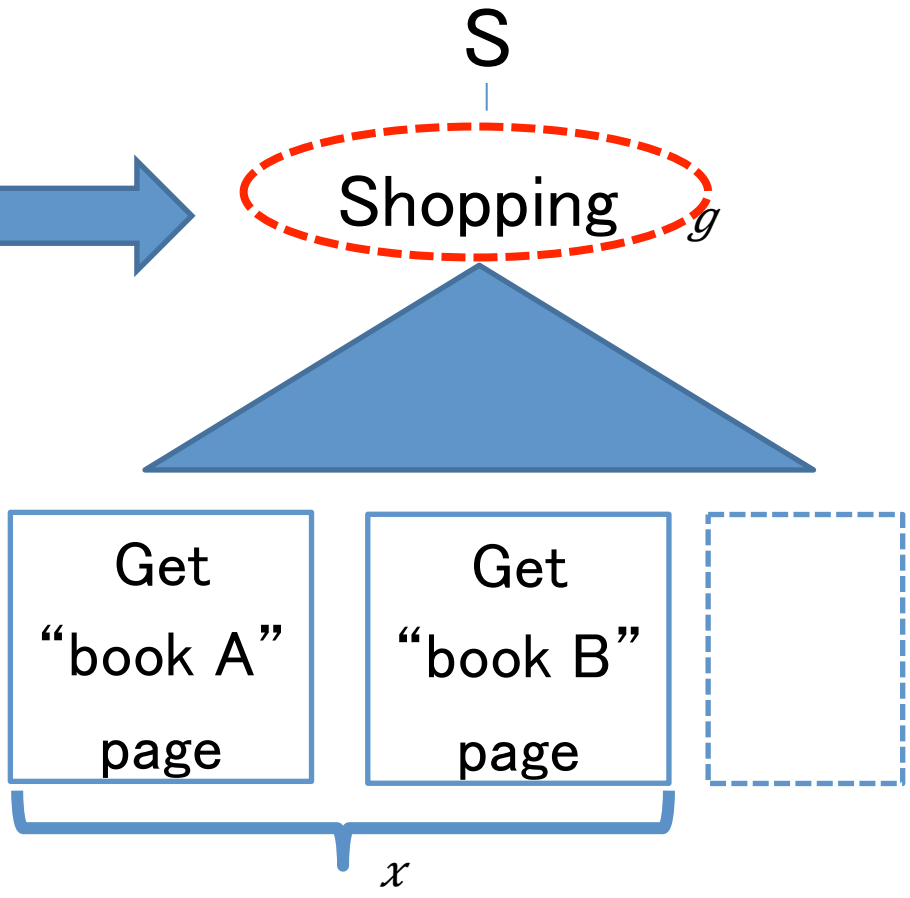
提案法の利用(1/2)

最尤構文木の出力 [Jelinek 85]
→　Webサイトの改善に利用



提案法の利用(2/2)

トッププラン
（ユーザの主な目的）
の推定
→広告表示などに利用



推定の式

$\mathrm{argmax}_{\top g} Pr(x,g)$

x : prefix

g :トッププラン

$Pr(x,g)$: gを経由して
x が導出される prefix 確率

- prefix x が与えられた時
- $\mathrm{argmax}_{\top g} Pr(x,g)$
- g:トッププラン
- $Pr(x,g)$: gを通って prefix x が導出される構文木の確率の和