知的情報処理システム特論　第9回

二宮　崇

今日の講義の予定

- 品詞解析
  - HMM
  - HMMの解析
    - ビタビアルゴリズム

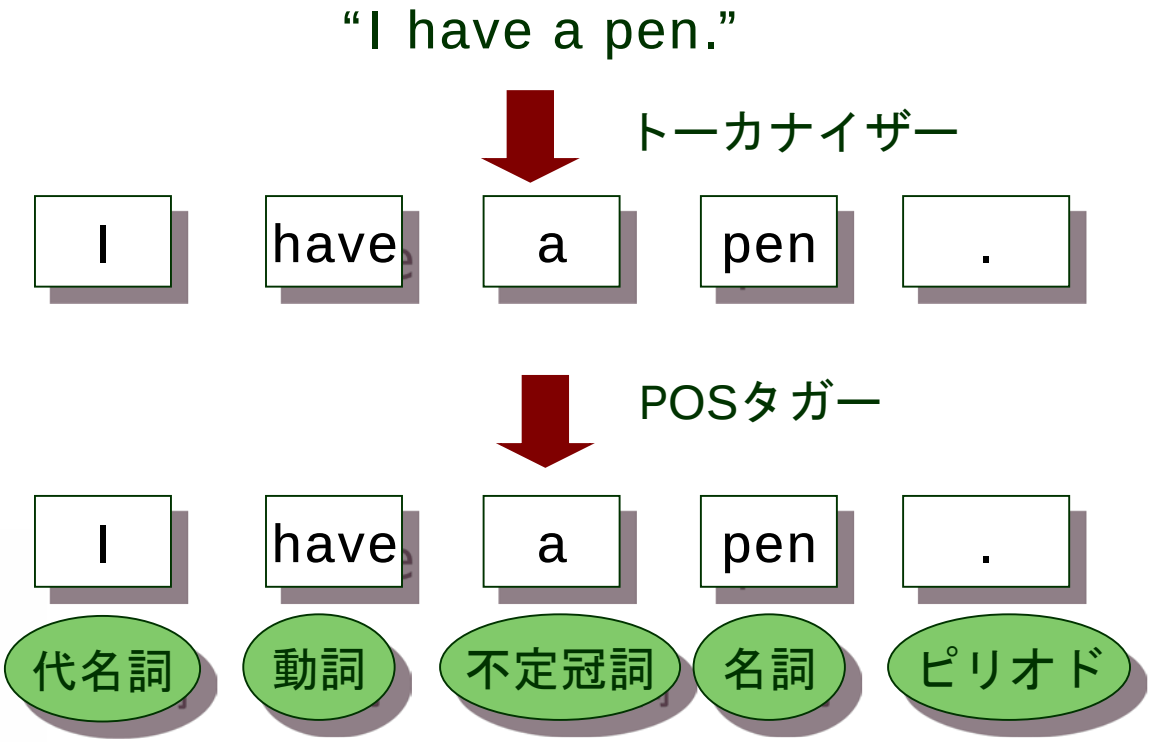
- 教科書
  - 北研二(著) 辻井潤一(編) 言語と計算4 確率的言語モデル 東大出版会
  - C. D. Manning & Hinrich Schütze “FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING” MIT Press, 1999
  - Christopher M. Bishop “PATTERN RECOGNITION AND MACHINE LEARNING” Springer, 2006

品詞解析

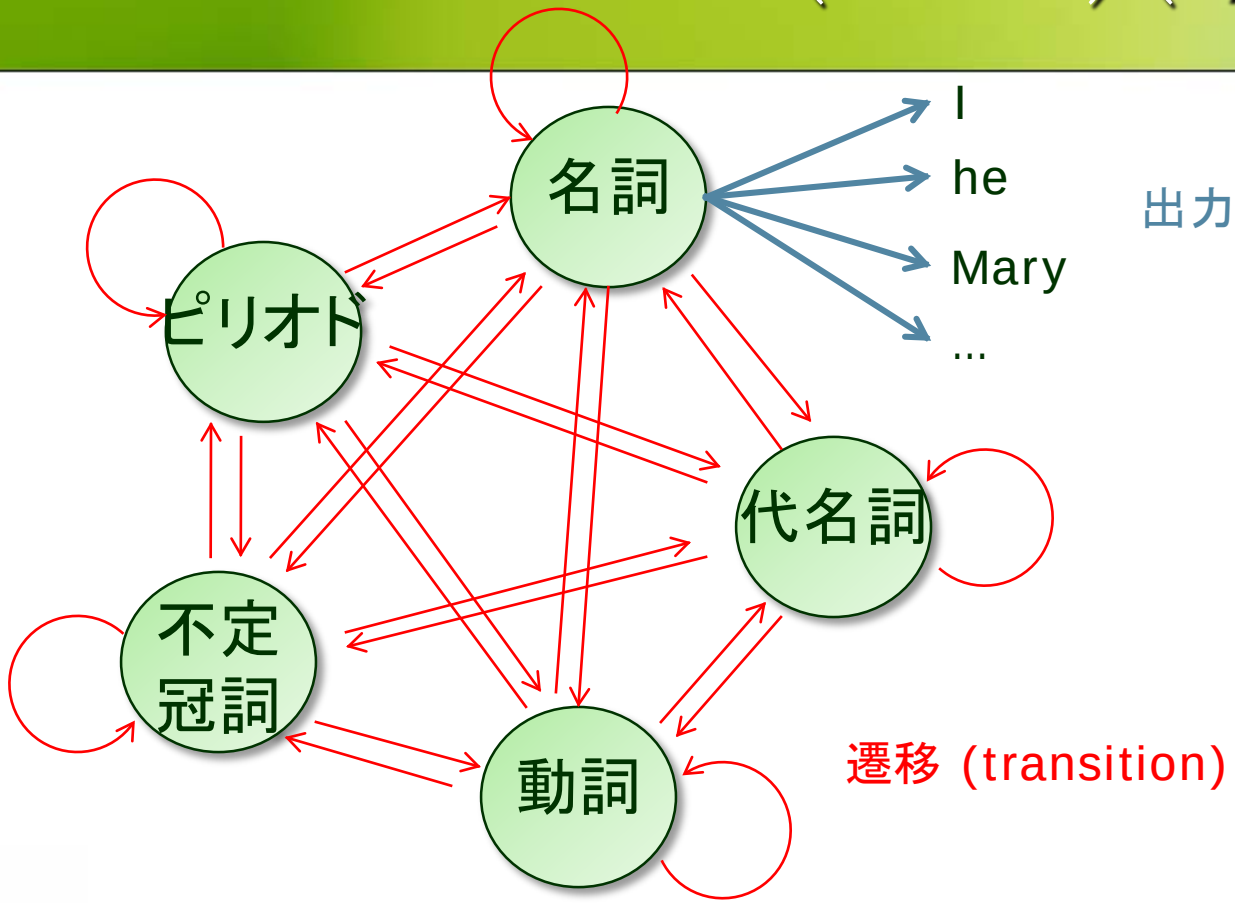
PART-OF-SPEECH (POS) TAGGING

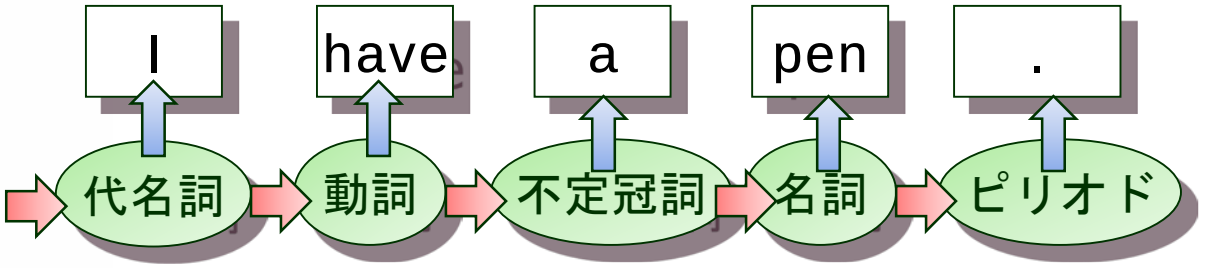
品詞解析

- 品詞タガー

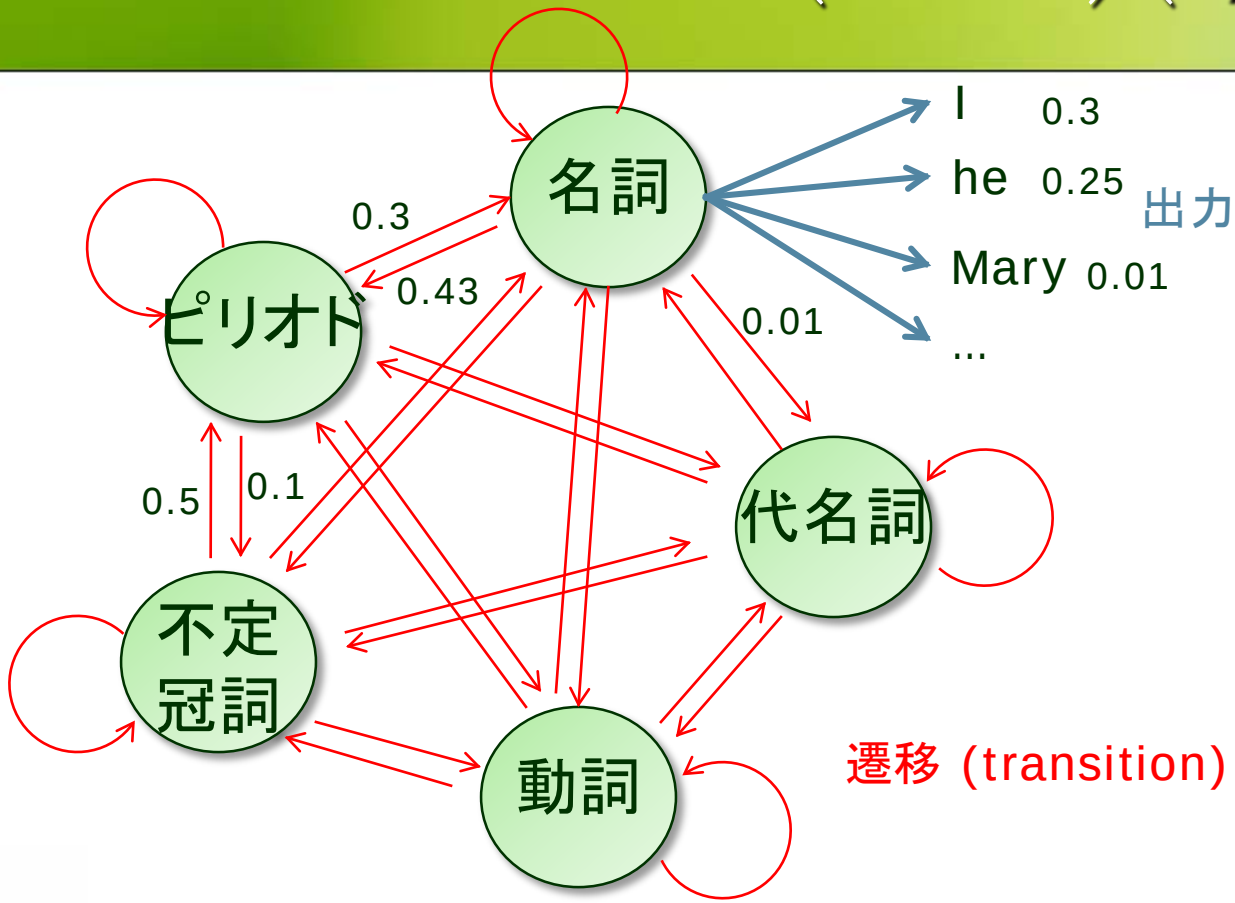

“I have a pen.”
トーカナイザー
I
have
a
pen
.
POSタガー
I
have
a
pen
.
代名詞
動詞
不定冠詞
名詞
ピリオド

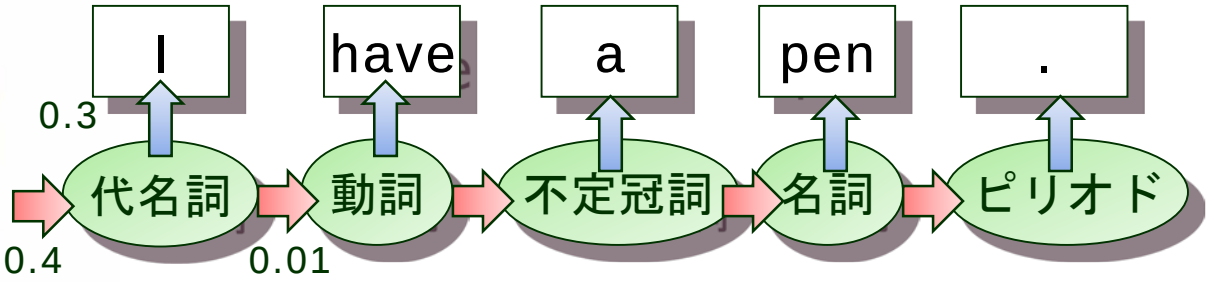
隠れマルコフモデル
Hidden Markov Model (HMM) (1/3)





隠れマルコフモデル
Hidden Markov Model (HMM) (2/3)





隠れマルコフモデル
Hidden Markov Model (HMM) (3/3)

- Q: 状態の有限集合
- Σ: 出力記号の有限集合
- π_q: 文頭が状態qになる確率
- $a_{q,r}$: 状態qから状態rへの遷移確率
  - $\Sigma_{r \in Q} a_{q,r} = 1$
- $b_{q,o}$: 状態qにおける記号oの出力確率
  - $\Sigma_{o \in \Sigma} b_{q,o} = 1$

状態記号列に対する確率の例

π

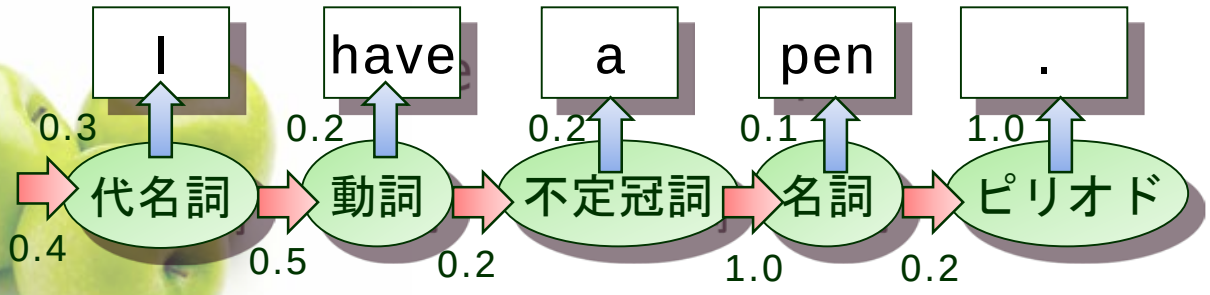
名詞	代名詞	動詞	不定冠詞	ピリオド
0.2	0.4	0.1	0.3	0.0

a

遷移元＼遷移先	名詞	代名詞	動詞	不定冠詞	ピリオド
名詞	0.5	0.0	0.3	0.0	0.2
代名詞	0.1	0.0	0.5	0.2	0.2
動詞	0.2	0.2	0.2	0.2	0.2
不定冠詞	1.0	0.0	0.0	0.0	0.0
ピリオド	0.5	0.0	0.5	0.0	0.0

b

状態＼出力	a	…	have	…	I	…	pen	…	.	…
名詞	0.01	…	0.0	…	0.01	…	0.1	…	0.0	…
代名詞	0.0	…	0.0	…	0.3	…	0.0	…	0.0	…
動詞	0.0	…	0.2	…	0.0	…	0.0	…	0.0	…
不定冠詞	0.2	…	0.0	…	0.0	…	0.0	…	0.0	…
ピリオド	0.0	…	0.0	…	0.0	…	0.0	…	1.0	…



0.4*0.3*0.5*0.2*0.2*
0.2*1.0*0.1*0.2*1.0
=0.0000096

状態記号列の確率と生成確率

- 状態と記号の列が与えられた時

状態記号列：$q_1 o_1 q_2 o_2 \cdots q_T o_T$

$$
\begin{aligned}
p(q_1 o_1 q_2 o_2 \cdots q_T o_T) &= \pi_{q_1} b_{q_1,o_1} a_{q_1,q_2} b_{q_2,o_2} \cdots a_{q_{T-1},q_T} b_{q_T,o_T} \\
&= \pi_{q_1} a_{q_1,q_2} \cdots a_{q_{T-1},q_T} b_{q_1,o_1} b_{q_2,o_2} \cdots b_{q_T,o_T} \\
&= \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1},q_t} \prod_{t=1}^{T} b_{q_t,o_t}
\end{aligned}
$$

- 記号列のみが与えられた時

記号列：$o_1 o_2 \cdots o_T$

$$
p(o_1 o_2 \cdots o_T) = \sum_{q_1 \in Q, q_2 \in Q, \cdots q_T \in Q} p(q_1 o_1 q_2 o_2 \cdots q_T o_T)
$$

（生成確率）

品詞解析

- 品詞解析 (入力: $o_1 o_2 \ldots o_T$)

$$\tilde{q}_1 \tilde{q}_2 \cdots \tilde{q}_T = \underset{q_1 \in Q, q_2 \in Q, \cdots q_T \in Q}{\arg\max} \; p(q_1 o_1 q_2 o_2 \cdots q_T o_T)$$

HMMの教師無し学習
Unsupervised Learning of HMMs

- パラメータ推定
  - 訓練データ (入力)

    訓練データ : $x_1 x_2 \cdots x_n$

    x_i : 文を表す記号列(単語列)。 $x_i = o_{i1} o_{i2} o_{i3} \cdots o_{iT_i}$とする。

    T_i : x_iの記号列長

  - パラメータ (出力)

$$\pi, a, b = \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(x_i)$$

$$= \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(o_{i1} o_{i2} \cdots o_{iT_i})$$

$$= \arg\max_{\pi,a,b} \prod_{i=1}^{n} \sum_{q_1 \in Q, q_2 \in Q, \cdots q_{T_i} \in Q} p(q_1 o_{i1} q_2 o_{i2} \cdots q_{T_i} o_{iT_i})$$

HMMの教師付学習
Supervised Learning of HMMs

- パラメータ推定
  - 訓練データ (入力)

訓練データ: $x_1 x_2 \cdots x_n, y_1 y_2 \cdots y_n$

x_i : 文を表す記号列(単語列)。 $x_i = o_{i1} o_{i2} o_{i3} \cdots o_{iT_i}$とする。

T_i : x_iの記号列長

y_i : x_iに対応する正解状態列(正解品詞列)。 $y_i = q_{i1} q_{i2} q_{i3} \cdots q_{iT_i}$とする。

  - パラメータ (出力)

$$\pi, a, b = \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(x_i, y_i)$$

$$= \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(q_{i1} o_{i1} q_{i2} o_{i2} \cdots q_{iT_i} o_{iT_i})$$

推論(INFERENCE)
解析(ANALYSIS)
タグ付け(TAGGING)
復号化(DECODING)

解析

- 解析 (入力: $o_1 o_2 \ldots o_T$)

$$\tilde{q}_1 \tilde{q}_2 \cdots \tilde{q}_T = \underset{q_1 \in Q, q_2 \in Q, \cdots q_T \in Q}{\arg\max} \; p(q_1 o_1 q_2 o_2 \cdots q_T o_T)$$

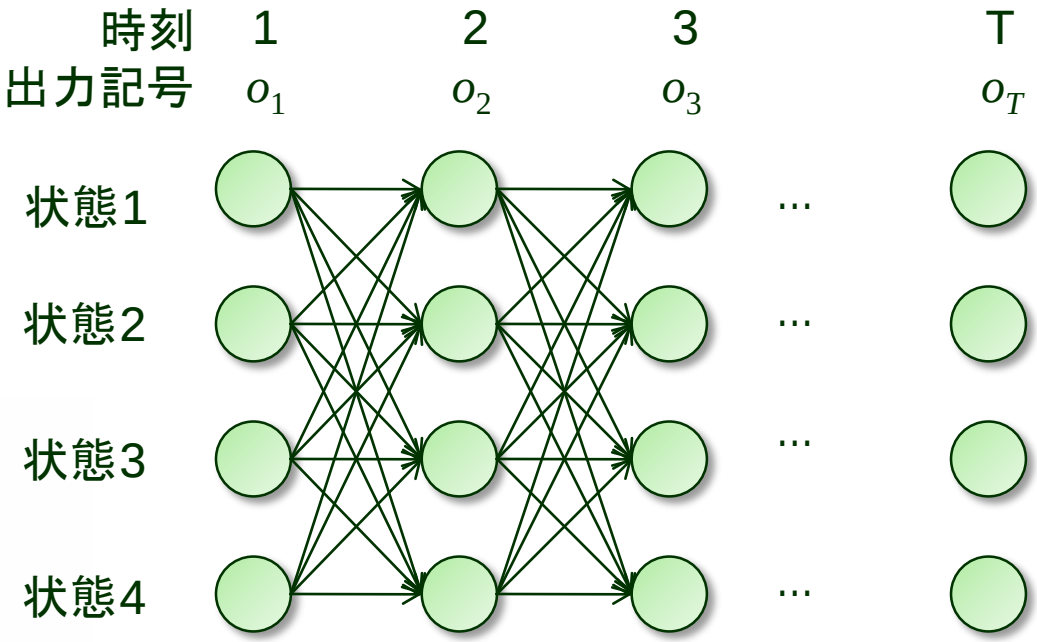
しかし、計算量は$O(|Q|^T)$!

効率的な品詞解析: ビタビアルゴリズム

- 動的計画法

$\delta(t, q)$: 時刻t(o_tが出力される時)に状態qとなる状態列の中で最大確率となる列の確率

$max_{q \in Q}\ \delta(T, q)$を求めれば良い

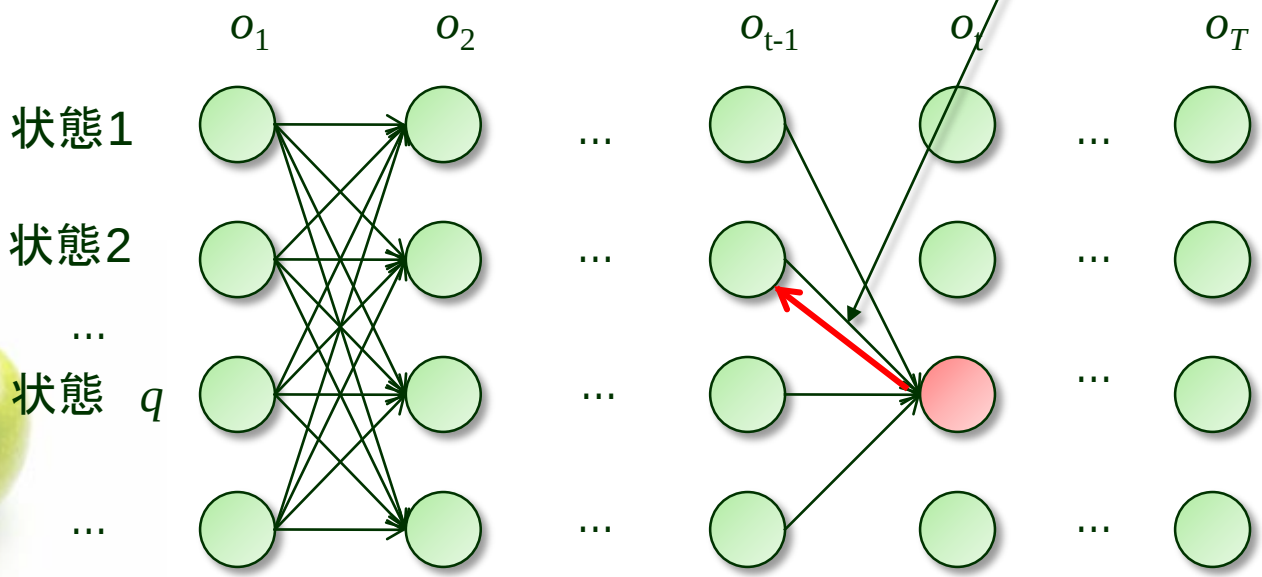


トレリス

効率的な品詞解析: ビタビアルゴリズム

$$
\begin{aligned}
\delta(t,q) &= \max_{q_1 \in Q,\cdots,q_{t-1} \in Q} p(q_1 o_1 \cdots q_{t-1} o_{t-1}) a_{q_{t-1},q} b_{q,o_t} \\
&= \max_{q_{t-1} \in Q} \left\{ \max_{q_1 \in Q,\cdots,q_{t-2} \in Q} \left\{ p(q_1 o_1 \cdots q_{t-2} o_{t-2}) a_{q_{t-2},q_{t-1}} b_{q_{t-1},o_{t-1}} \right\} a_{q_{t-1},q} b_{q,o_t} \right\} \\
&= \max_{q_{t-1} \in Q} \delta(t-1, q_{t-1}) a_{q_{t-1},q} b_{q,o_t}
\end{aligned}
$$

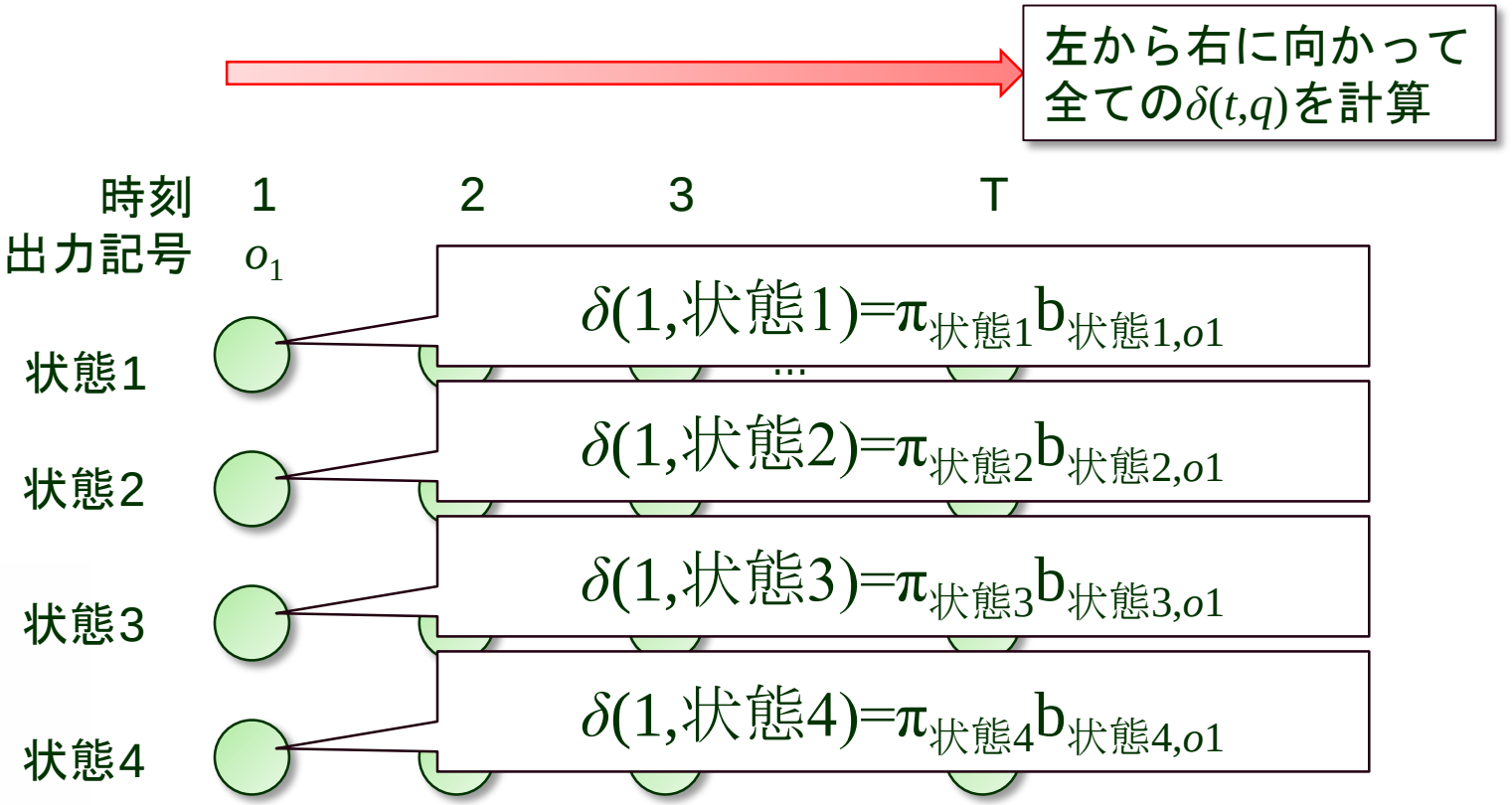
最大確率で遷移したパスをバックポインタで保存



効率的な品詞解析: ビタビアルゴリズム

- 動的計画法

$\delta(t, q)$: 時刻t(o_tが出力される時)に状態qとなる状態列の中で最大確率となる列の確率



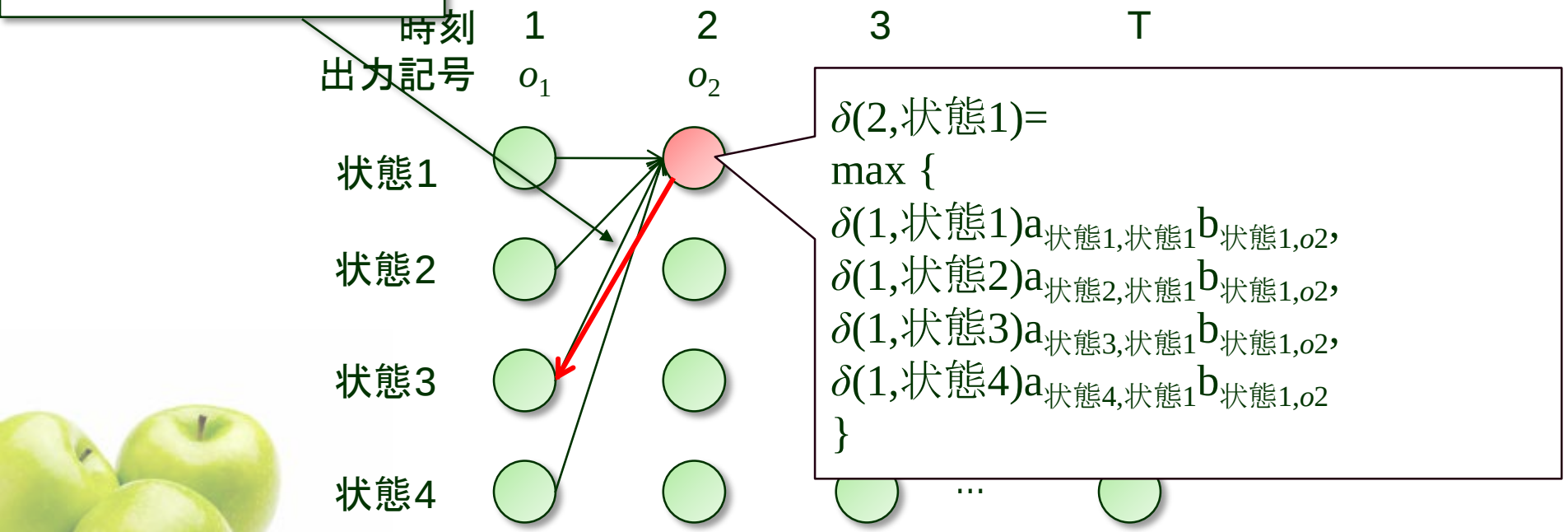
効率的な品詞解析: ビタビアルゴリズム

- 動的計画法

$\delta(t, q)$: 時刻t(o_tが出力される時)に状態qとなる状態列の中で最大確率となる列の確率

最大確率で遷移したパスをバックポインタで保存

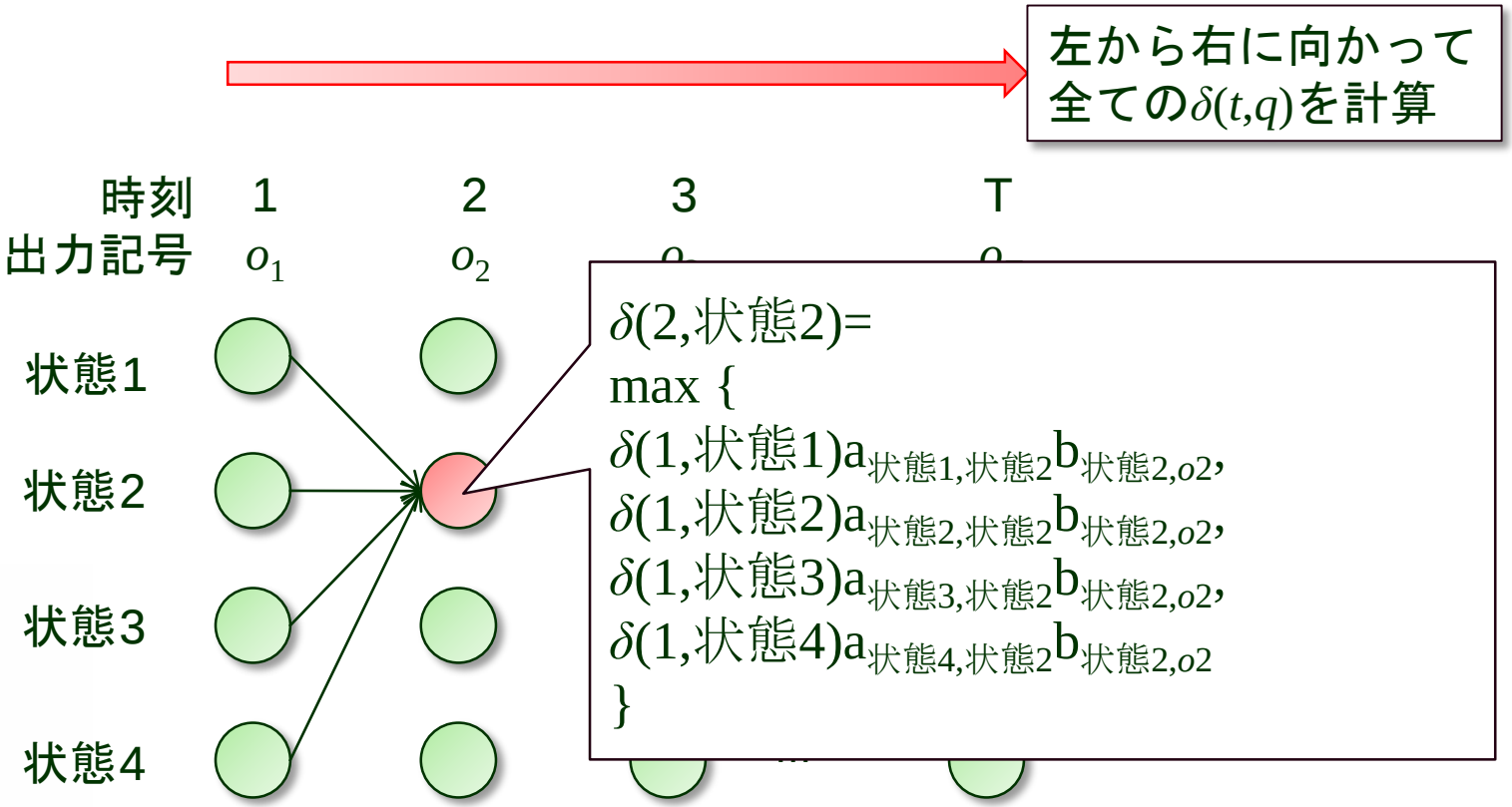
左から右に向かって全ての$\delta(t,q)$を計算



効率的な品詞解析: ビタビアルゴリズム

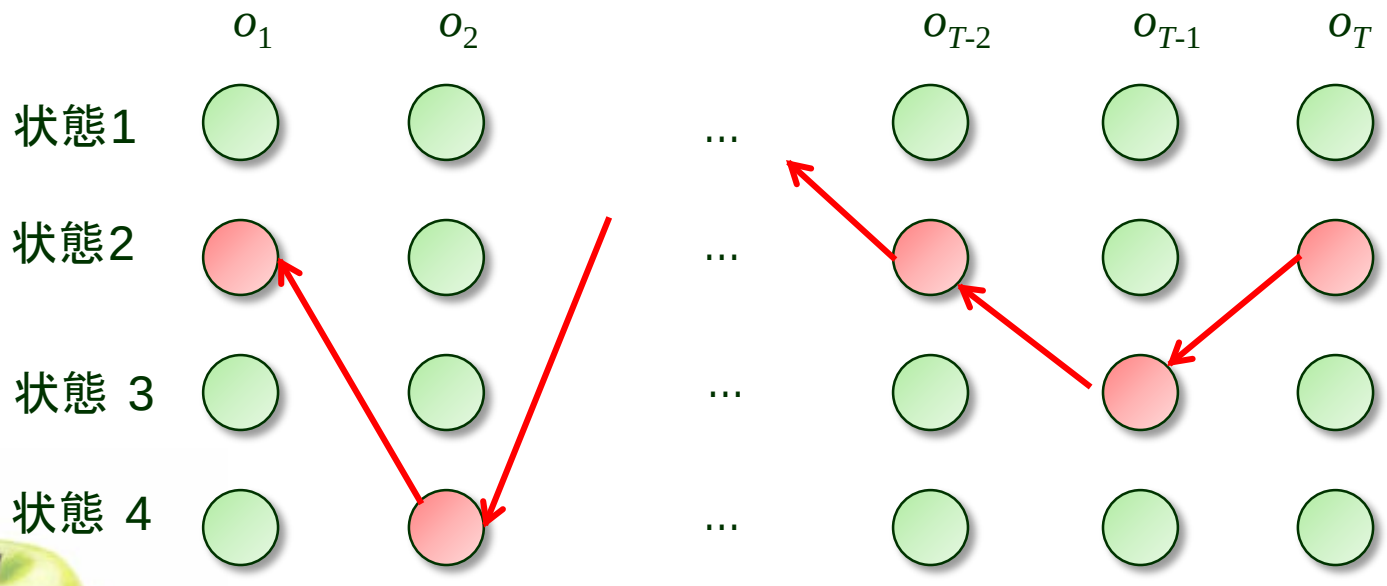
- 動的計画法

$\delta(t, q)$: 時刻t(o_tが出力される時)に状態qとなる状態列の中で最大確率となる列の確率



効率的な品詞解析: ビタビアルゴリズム

- 最後にバックポインタを辿ることで最大確率となる状態列が得られる



効率的な品詞解析: ビタビアルゴリズム

$\delta[1,q] := \pi[q]b[q, o_1]$ (for all q)

for t =2 to T

  for $q \in Q$

    $\delta[t, q] := \max_{q' \in Q} \{\delta[t-1, q']a[q',q]b[q, o_t]\}$

    $bp[t,q] := \text{argmax}_{q' \in Q} \{\delta[t-1, q']a[q',q]b[q, o_t]\}$

まとめ

- 品詞解析
  - HMM
  - HMMの解析
    - ビタビアルゴリズム
- 資料

  http://aiweb.cs.ehime-u.ac.jp/~ninomiya/iips/