人工知能特論II　第6回

二宮　崇

今日の講義の予定

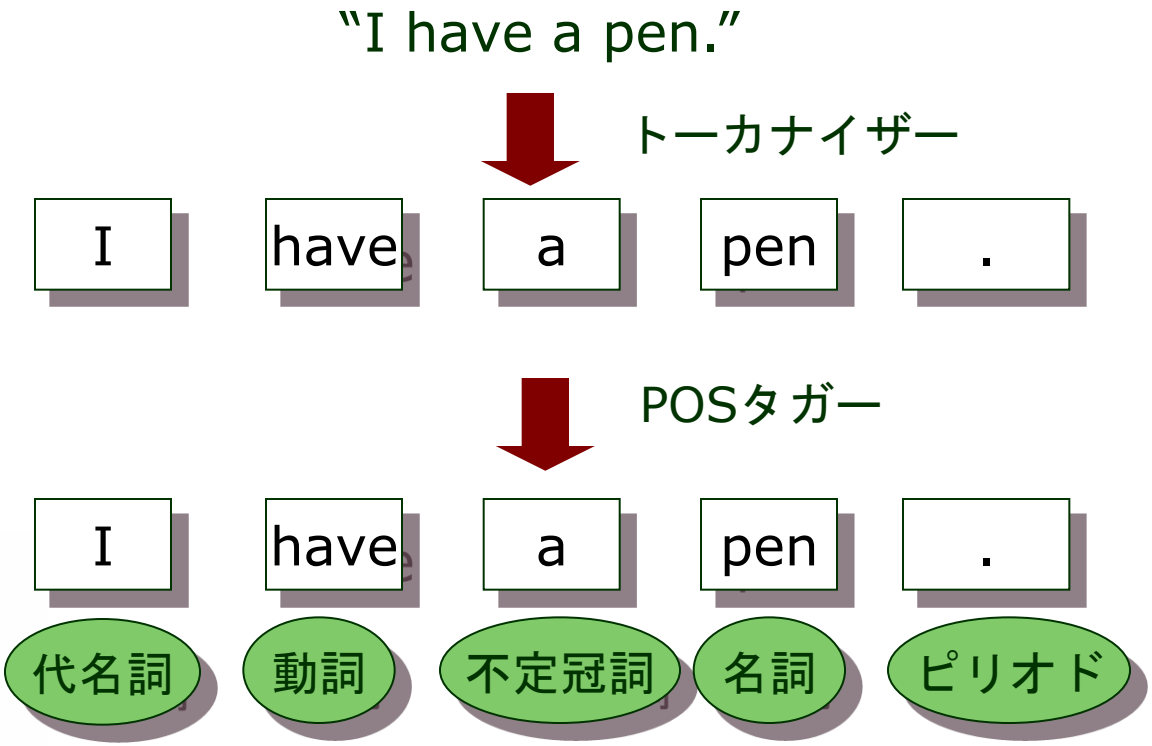
- 確率的文法
  - 品詞解析
    - HMM
  - 構文解析
    - PCFG
- 教科書
  - 北研二(著) 辻井潤一(編) 言語と計算4 確率的言語モデル 東大出版会
  - C. D. Manning & Hinrich Schütze “FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING” MIT Press, 1999
  - Christopher M. Bishop “PATTERN RECOGNITION AND MACHINE LEARNING” Springer, 2006

品詞解析

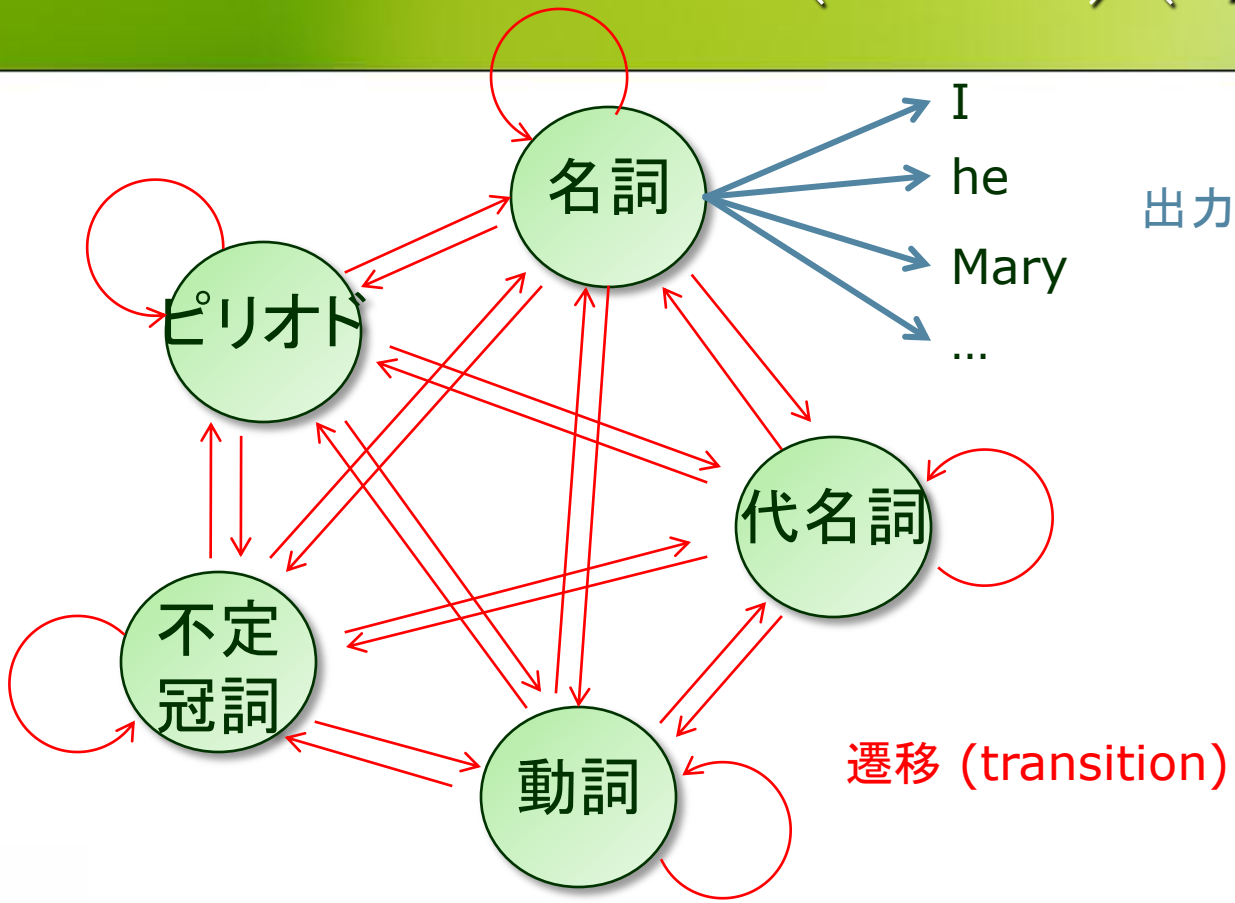
PART-OF-SPEECH (POS) TAGGING

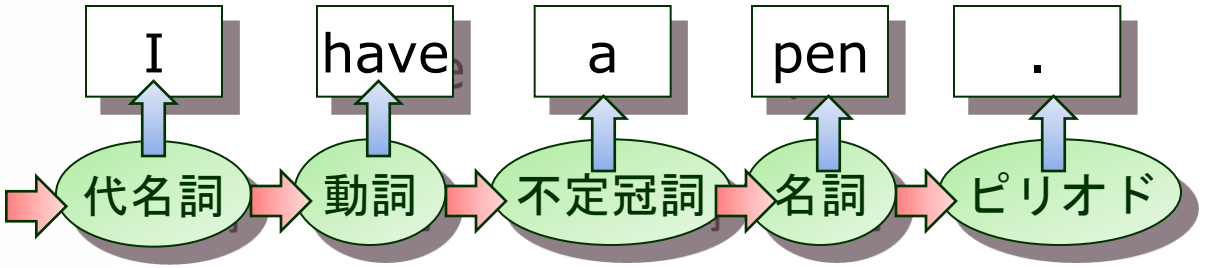
品詞解析

- 品詞タガー

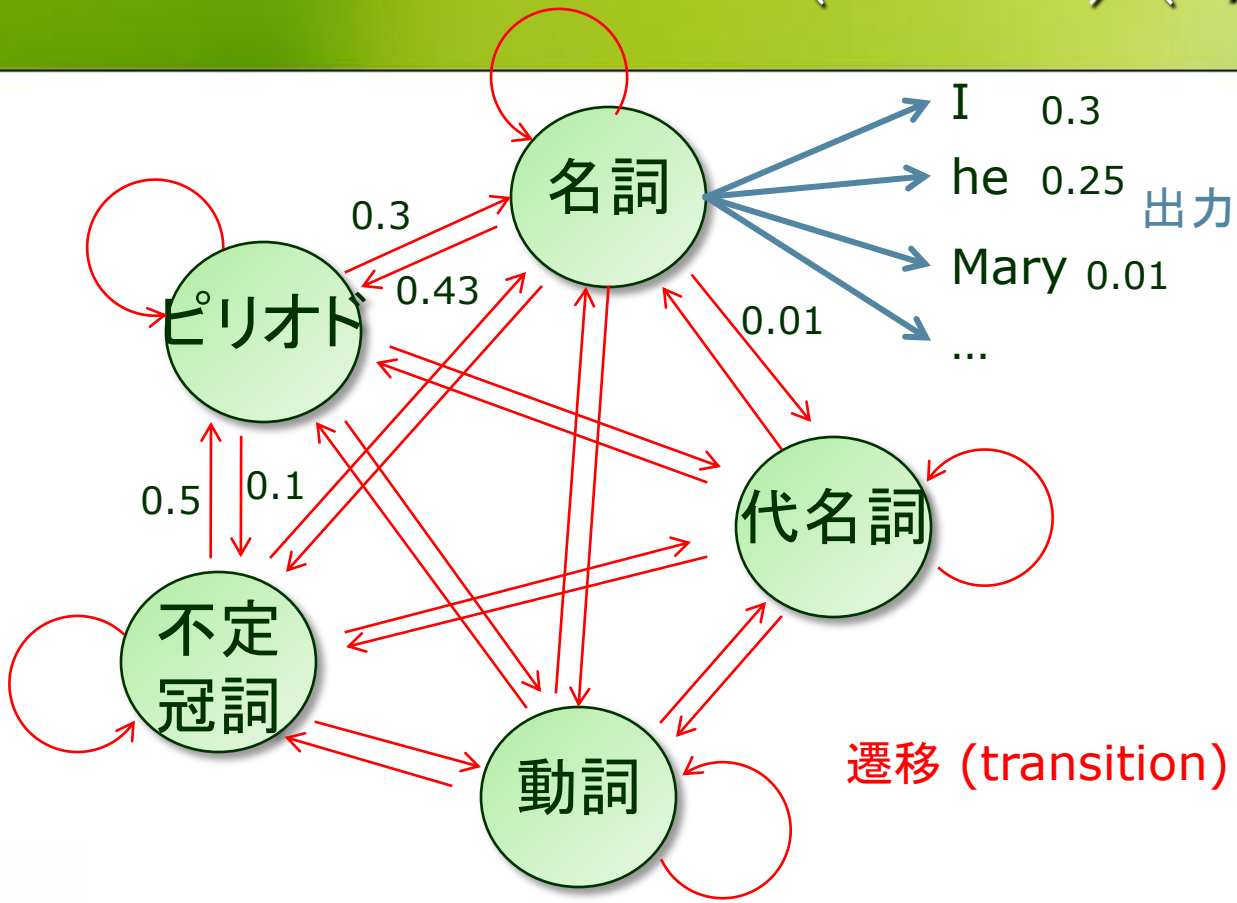

“I have a pen.”
トーカナイザー
I
have
a
pen
.
POSタガー
I
have
a
pen
.
代名詞
動詞
不定冠詞
名詞
ピリオド

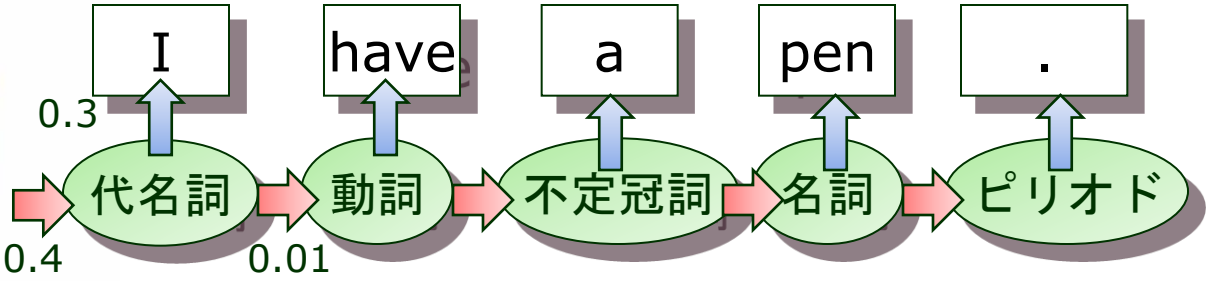
隠れマルコフモデル
Hidden Markov Model (HMM) (1/3)





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

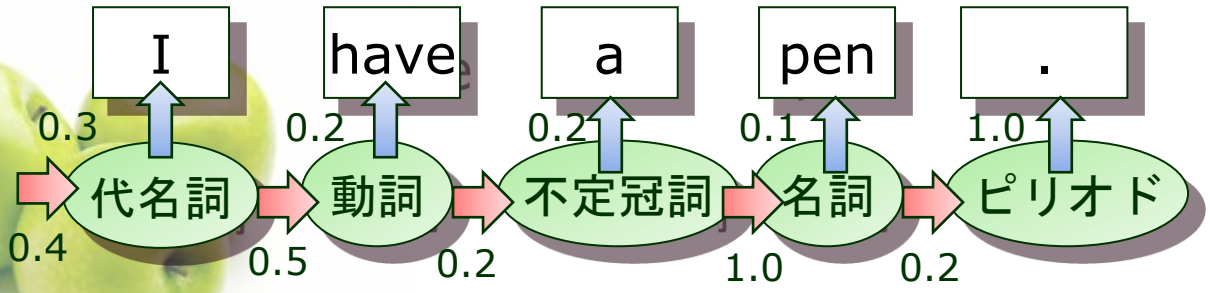
名詞	代名詞	動詞	不定冠詞	ピリオド
0.2	0.4	0.1	0.3	0.0

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)

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

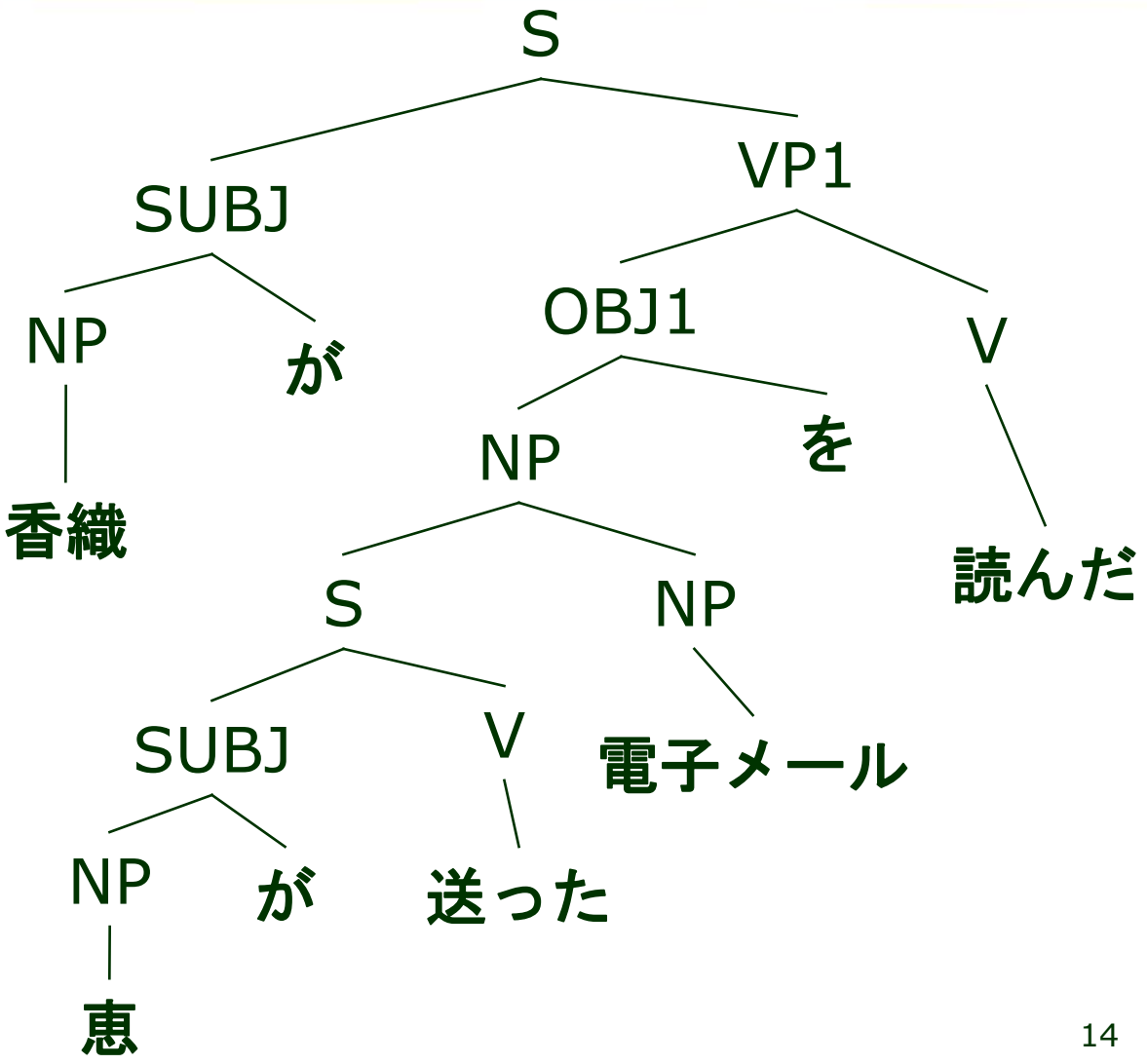
$$= \arg\max_{\pi,a,b} \prod_{i=1}^{n} p(q_{i1} o_{i1} q_{i2} o_{i2} \cdots q_{iT_i} o_{iT_i})$$

確率的文脈自由文法 (PCFG)

PROBABILISTIC CONTEXT FREE GRAMMAR (PCFG)

CFG: 構文木

簡単なCFGの例
S → SUBJ　VP1
S → SUBJ　V
SUBJ → NP　が
VP1 → OBJ1　V
OBJ1 → NP　を
NP → S　NP
V → 送った
V → 読んだ
NP → 香織
NP → 恵
NP → 電子メール
NP → プレゼント
NP → 香織　NP1
NP → 恵　NP1
NP1 → と　NP



確率的CFG (PCFG)

- CFGの書換規則の適用確率をパラメータ化した文法
- 構文木の確率は、適用された書換規則のパラメータの積
- 各パラメータは$0.0 \leqq \theta \leqq 1.0$

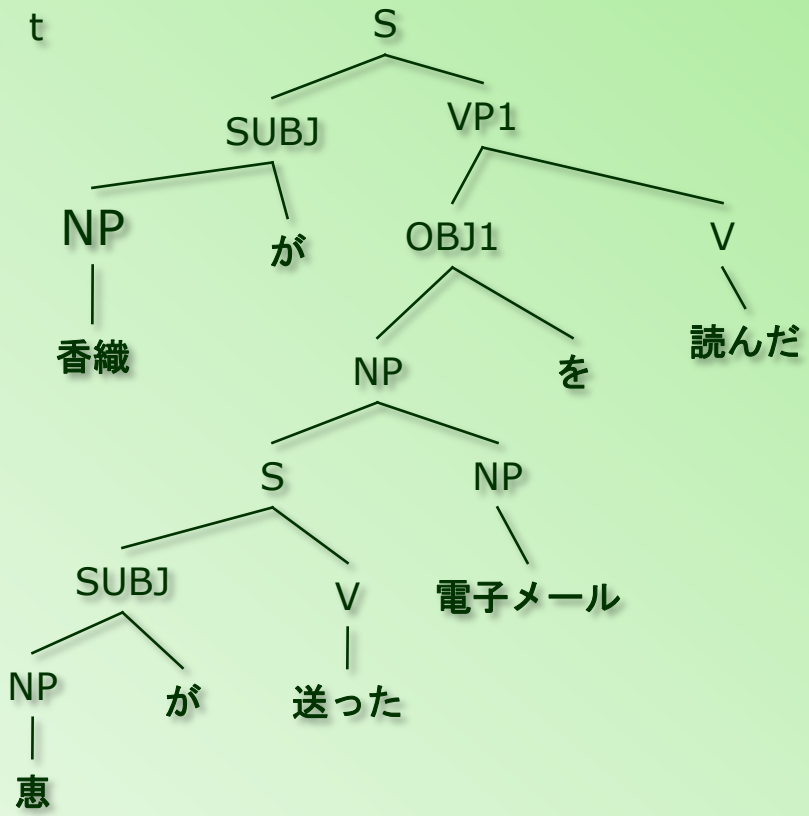
簡単なCFGの例	パラメータ
S → SUBJ　VP1	$\theta_{S \rightarrow SUBJ\ \ VP1}$
S → SUBJ　V	$\theta_{S \rightarrow SUBJ\ \ V}$
SUBJ → NP　が	$\theta_{SUBJ \rightarrow NP\ \ が}$
VP1 → OBJ1　V	$\theta_{VP1 \rightarrow OBJ1\ \ V}$
OBJ1 → NP　を	$\theta_{OBJ1 \rightarrow NP\ を}$
NP → S　NP	$\theta_{NP \rightarrow S\ \ NP}$
V → 送った	$\theta_{V \rightarrow 送った}$
V → 読んだ	$\theta_{V \rightarrow 読んだ}$
NP → 香織	$\theta_{NP \rightarrow 香織}$
NP → 恵	$\theta_{NP \rightarrow 恵}$
NP → 電子メール	$\theta_{NP \rightarrow 電子メール}$
NP → プレゼント	$\theta_{NP \rightarrow プレゼント}$
NP → 香織　NP1	$\theta_{NP \rightarrow 香織\ NP1}$
NP → 恵　NP1	$\theta_{NP \rightarrow 恵\ \ NP1}$
NP1 → と　NP	$\theta_{NP1 \rightarrow と\ \ NP}$

構文木の確率

簡単なCFGの例	パラメータ
S → SUBJ VP1	$\theta_{S \rightarrow SUBJ\ VP1}$
S → SUBJ V	$\theta_{S \rightarrow SUBJ\ V}$
SUBJ → NP が	$\theta_{SUBJ \rightarrow NP\ が}$
VP1 → OBJ1 V	$\theta_{VP1 \rightarrow OBJ1\ V}$
OBJ1 → NP を	$\theta_{OBJ1 \rightarrow NP\ を}$
NP → S NP	$\theta_{NP \rightarrow S\ NP}$
V → 送った	$\theta_{V \rightarrow 送った}$
V → 読んだ	$\theta_{V \rightarrow 読んだ}$
NP → 香織	$\theta_{NP \rightarrow 香織}$
NP → 恵	$\theta_{NP \rightarrow 恵}$
NP → 電子メール	$\theta_{NP \rightarrow 電子メール}$
NP → プレゼント	$\theta_{NP \rightarrow プレゼント}$
NP → 香織 NP1	$\theta_{NP \rightarrow 香織\ NP1}$
NP → 恵 NP1	$\theta_{NP \rightarrow 恵\ NP1}$
NP1 → と NP	$\theta_{NP1 \rightarrow と\ NP}$

文 x = “香織が恵が送った電子メールを読んだ”

構文木 t



$$P(t) = \theta_{S \rightarrow SUBJ\ VP1} \times \theta_{SUBJ \rightarrow NP\ が} \times \theta_{NP \rightarrow 香織} \times \theta_{VP1 \rightarrow OBJ1\ V} \times \theta_{OBJ1 \rightarrow NP\ を} \times \theta_{NP \rightarrow S\ NP} \times \theta_{S \rightarrow SUBJ\ V} \times \theta_{SUBJ \rightarrow NP\ が} \times \theta_{NP \rightarrow 恵} \times \theta_{V \rightarrow 送った} \times \theta_{NP \rightarrow 電子メール} \times \theta_{V \rightarrow 読んだ}$$

書換規則の制約

- Sからの生成
  - S→SUBJ　VP1を使う場合
  - S→SUBJ　Vを使う場合
- Sからの生成は2通りしかないので、
  - $\theta_{S \rightarrow SUBJ\ \ VP1} + \theta_{S \rightarrow SUBJ\ \ V} = 1.0$
- 同様にNPも
  - $\theta_{NP \rightarrow S\ \ NP} + \theta_{NP \rightarrow 香織} + \theta_{NP \rightarrow 恵} + \theta_{NP \rightarrow 電子メール} + \theta_{NP \rightarrow プレゼント} + \theta_{NP \rightarrow 香織\ NP1} + \theta_{NP \rightarrow 恵\ \ NP1} = 1.0$

簡単なCFGの例	パラメータ
S → SUBJ　VP1	$\theta_{S \rightarrow SUBJ\ \ VP1}$
S → SUBJ　V	$\theta_{S \rightarrow SUBJ\ \ V}$
SUBJ → NP　が	$\theta_{SUBJ \rightarrow NP\ \ が}$
VP1 → OBJ1　V	$\theta_{VP1 \rightarrow OBJ1\ \ V}$
OBJ1 → NP　を	$\theta_{OBJ1 \rightarrow NP\ を}$
NP → S　NP	$\theta_{NP \rightarrow S\ \ NP}$
V → 送った	$\theta_{V \rightarrow 送った}$
V → 読んだ	$\theta_{V \rightarrow 読んだ}$
NP → 香織	$\theta_{NP \rightarrow 香織}$
NP → 恵	$\theta_{NP \rightarrow 恵}$
NP → 電子メール	$\theta_{NP \rightarrow 電子メール}$
NP → プレゼント	$\theta_{NP \rightarrow プレゼント}$
NP → 香織　NP1	$\theta_{NP \rightarrow 香織\ NP1}$
NP → 恵　NP1	$\theta_{NP \rightarrow 恵\ \ NP1}$
NP1 → と　NP	$\theta_{NP1 \rightarrow と\ \ NP}$

書換規則の制約

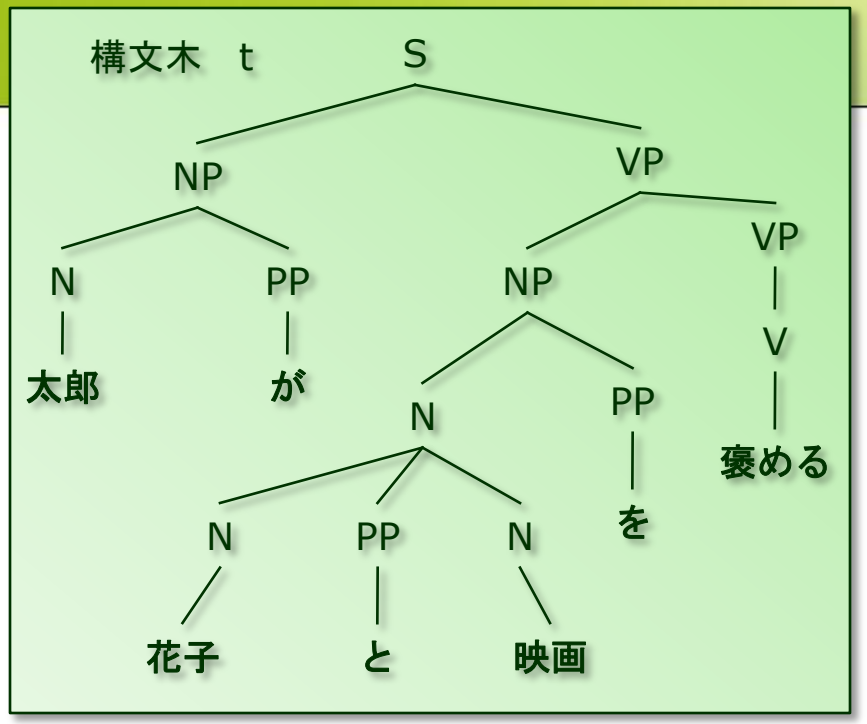
- CFG書換規則をA→ αと表したとき、(Aは非終端記号、αは非終端記号列)すべての非終端記号Aに対し、

$$\sum_{\alpha} \theta_{A \to \alpha} = 1.0$$

とする。

構文木の確率

CFG G	パラメータ	値
S → NP VP	$\theta_{S \to NP\ VP}$	1.0
NP → N PP	$\theta_{NP \to N\ PP}$	1.0
N → N PP N	$\theta_{N \to N\ PP\ N}$	0.1
VP → NP VP	$\theta_{VP \to NP\ VP}$	0.3
VP → V	$\theta_{VP \to V}$	0.7
PP → が	$\theta_{PP \to が}$	0.5
PP → を	$\theta_{PP \to を}$	0.3
PP → と	$\theta_{PP \to と}$	0.2
N → 太郎	$\theta_{N \to 太郎}$	0.3
N → 花子	$\theta_{N \to 花子}$	0.2
N → 映画	$\theta_{N \to 映画}$	0.4
V → 褒める	$\theta_{V \to 褒める}$	0.3
V → 見る	$\theta_{V \to 見る}$	0.7



$$p(t) = \theta_{S \to NP\ VP} \times \theta_{NP \to N\ PP} \times \theta_{N \to 太郎} \times \theta_{PP \to が} \times \theta_{VP \to NP\ VP} \times \theta_{NP \to N\ PP} \times \theta_{N \to N\ PP\ N} \times \theta_{N \to 花子} \times \theta_{PP \to と} \times \theta_{N \to 映画} \times \theta_{PP \to を} \times \theta_{VP \to V} \times \theta_{V \to 褒める}$$

$$= 1.0 \times 1.0 \times 0.3 \times 0.5 \times 0.3 \times 1.0 \times 0.1 \times 0.2 \times 0.2 \times 0.4 \times 0.3 \times 0.7 \times 0.3$$

$$= 0.000004536$$

構文木tの確率

- C(r; t): CFG<V_N, V_T, P, σ>の書換規則$r \in P$が構文木t中で使われた回数

$$p(t) = \prod_{r \in P} \theta_r^{C(r;t)}$$

文の生成確率

- ある文xに対し、xを導出する全ての構文木集合を$T(x)$としたとき、

$$p(x) = \sum_{t \in T(x)} p(t) \qquad 0.0 \leq p(x) \leq 1.0$$

- PCFGによって生成されうる全ての構文木集合をTとしたとき、

$$\sum_{t \in T} p(t) \approx 1.0$$

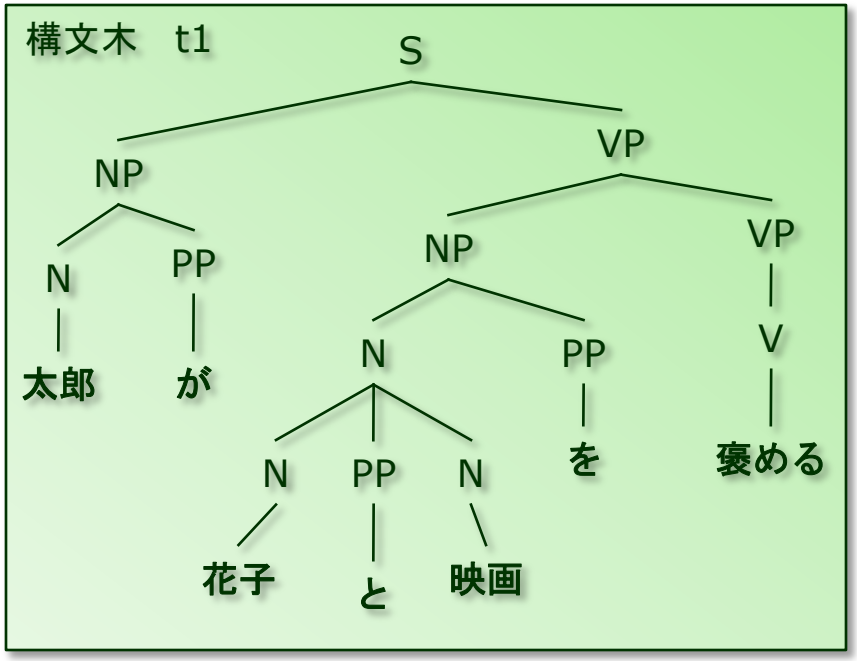
構文解析

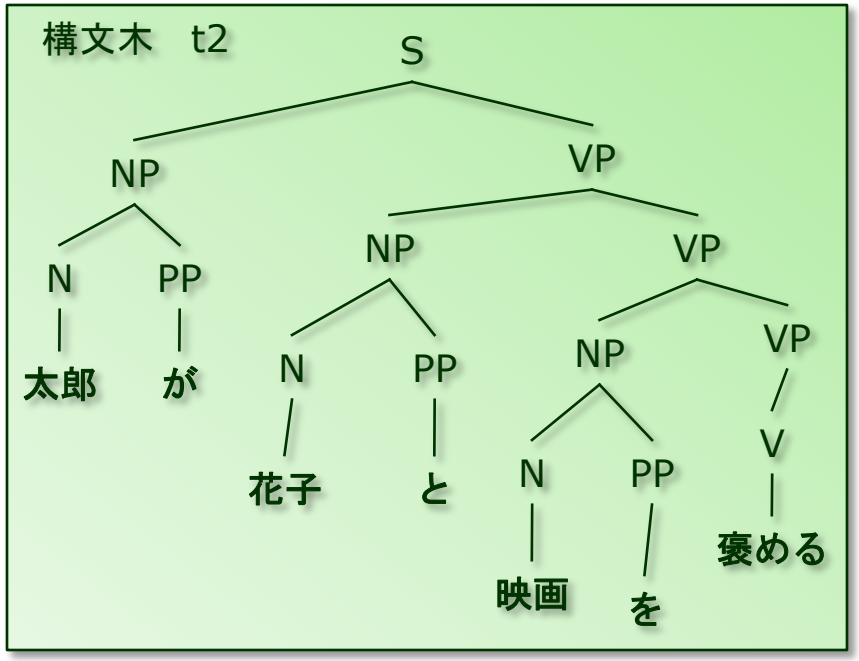
- ある文xに対し、CFG<V_N, V_T, P, σ>を用いてxを導出できる全ての構文木集合を$T(x)$としたとき、

$$\tilde{t} = \arg\max_{t \in T(x)} p(t)$$
$$= \arg\max_{t \in T(x)} \prod_{r \in P} \theta_r^{C(r;t)}$$

構文解析の例

- ”太郎が花子と映画を褒める”に対する構文木は次の二種類しかない





$p(t1) = \theta_{S \to NP\ VP} \times \theta_{NP \to N\ PP}^{2} \times \theta_{N \to 太郎} \times \theta_{PP \to が} \times \theta_{VP \to NP\ VP} \times \theta_{N \to N\ PP\ N} \times \theta_{N \to 花子} \times \theta_{PP \to と} \times \theta_{N \to 映画} \times \theta_{PP \to を} \times \theta_{VP \to V} \times \theta_{V \to 褒める}$

$= 1.0 \times 1.0^{2} \times 0.3 \times 0.5 \times 0.3 \times 0.1 \times 0.2 \times 0.2 \times 0.4 \times 0.3 \times 0.7 \times 0.3$

$= 0.000004536$

$p(t2) = \theta_{S \to NP\ VP} \times \theta_{NP \to N\ PP}^{3} \times \theta_{N \to 太郎} \times \theta_{PP \to が} \times \theta_{VP \to NP\ VP}^{2} \times \theta_{N \to 花子} \times \theta_{PP \to と} \times \theta_{N \to 映画} \times \theta_{PP \to を} \times \theta_{VP \to V} \times \theta_{V \to 褒める}$

$= 1.0 \times 1.0^{3} \times 0.3 \times 0.5 \times 0.3^{2} \times 0.2 \times 0.2 \times 0.4 \times 0.3 \times 0.7 \times 0.3$

$= 0.000013608$

PCFGの教師無し学習
Unsupervised Learning of PCFG

- 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

  - パラメータ (出力)

$$\begin{aligned}\boldsymbol{\theta} &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(x_i) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \sum_{t \in T(x_i)} p(t) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \sum_{t \in T(x_i)} \prod_{r \in P} \theta_r^{C(r;t)}\end{aligned}$$

PCFGの教師付パラメータ推定
Supervised Learning of PCFG

- 教師付学習
  - 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

    y_i : x_iに対する正解構文木

  - パラメータ (出力)

$$\begin{aligned}\boldsymbol{\theta} &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(x_i, y_i) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(y_i) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \prod_{r \in P} \theta_r^{C(r;y_i)}\end{aligned}$$

まとめ

- 品詞解析
  - HMM
- 構文解析
  - PCFG
- 資料

  http://aiweb.cs.ehime-u.ac.jp/~ninomiya/ai2/