Keio University SFC 2004

『探索的モデリング』

第11回 隠れマルコフモデル①

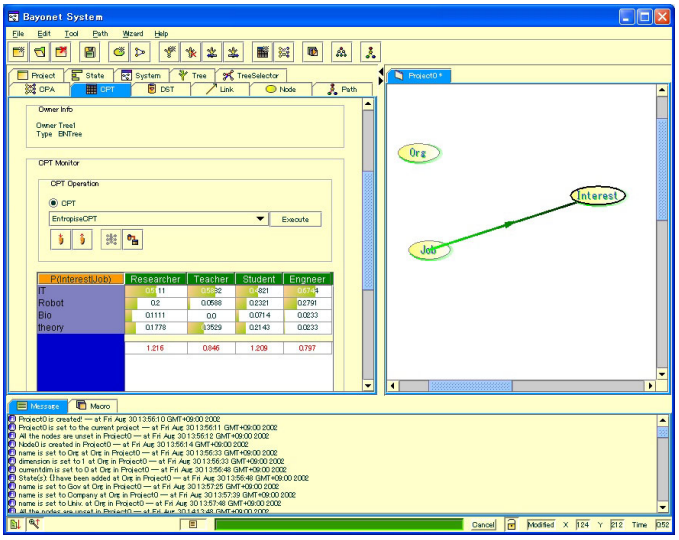
いば　たかし
井庭 崇
慶應義塾大学総合政策学部 専任講師
iba@sfc.keio.ac.jp
http://www.sfc.keio.ac.jp/~iba/lecture/

補講（2回分）

- １月２１日（金）
- ε２２教室

- ４限：隠れマルコフモデル②
- ５限：最終レポート相談会

ベイジアンネット構築ソフトウェア「BayoNet」



期末レポート

- どちらかのテーマを選択してレポートにまとめてください。

  【テーマ1】自分の関心のある分野に対して、ベイジアンネットワークまたは隠れマルコフモデルを適用し、解説してください。

  【テーマ2】先行研究のサーベイを行い、それらの研究について解説してください（Webで「Bayesian Network」で検索すると、たくさん論文が見つかります。必ずURLや文献名を記述してください）。

- レポートシステムで提出
  - レポートシステムの準備が整い次第、メールでお知らせします。
  - 提出期限：1月28日（金）
  - 本文：A4用紙で5枚以内
  - doc または pdf 形式： ファイル容量3 MBまで
  - 必ず参考文献を明記すること

Keio University SFC 2004

『探索的モデリング』

第１１回 隠れマルコフモデル①

いば　たかし
井庭 崇
慶應義塾大学総合政策学部 専任講師
iba@sfc.keio.ac.jp
http://www.sfc.keio.ac.jp/~iba/lecture/

隠れマルコフモデルとは？

隠れマルコフモデル

- 確率的な状態遷移と確率的な記号出力を備えたマルコフモデル（入力に対して状態を変えていくオートマトンと捉えてよい）
- 外部から観測できるのは記号出力の系列だけであり、内部の状態遷移は直接観測できないところから「隠れ（hidden）」マルコフモデルと呼ばれる。
- 隠れマルコフモデルは状態が確率的に遷移するので、同じ入力があっても、状態がAになる時もあるし、Bになる時もある。
- この確率を変えることで、学習させることができる。

隠れマルコフモデルの登場によって・・・

- 隠れマルコフモデルは、（観測可能な）言語データから言語現象の背後にある（隠れた）構造を推定する場合に有効なモデルである。
- 研究者が頭を悩ませなくても、最初にうまく設計しておけば良いということ
  - データはとにかく用意すればいい。結果と比較する手間も必要ない。
  - 隠れマルコフモデルの設計をうまくやっておけば、使ううちに良くなってくる（学習）。
- それをうまく商品にしたのが、ViaVoice。

隠れマルコフモデルの応用事例

- 隠れマルコフモデルは、音声認識のための音響モデルとして標準的に用いられている。
  - 1970年代にIBMの音声グループが音声認識に隠れマルコフモデルを適用し大きな成功を収めたことに端を発している
  - 1997年夏に登場した「Voice Type」や「Via Voice」などで、機械の音声認識率が飛躍的に増大したことで、隠れマルコフモデルは有名になった。
- 自然言語処理においても、確率的形態素解析を初めとするさまざまなところに適用されている。
- ゲノム解析、認識

隠れマルコフモデルの例①
英語の品詞判別

自然言語の確率・統計的性質

単語の出現頻度分布

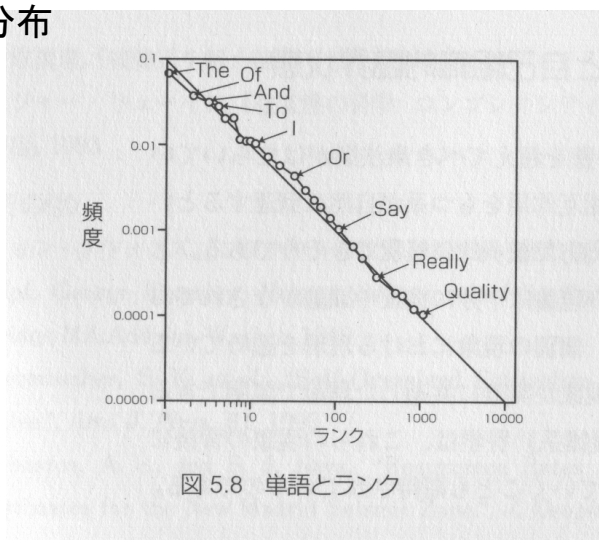


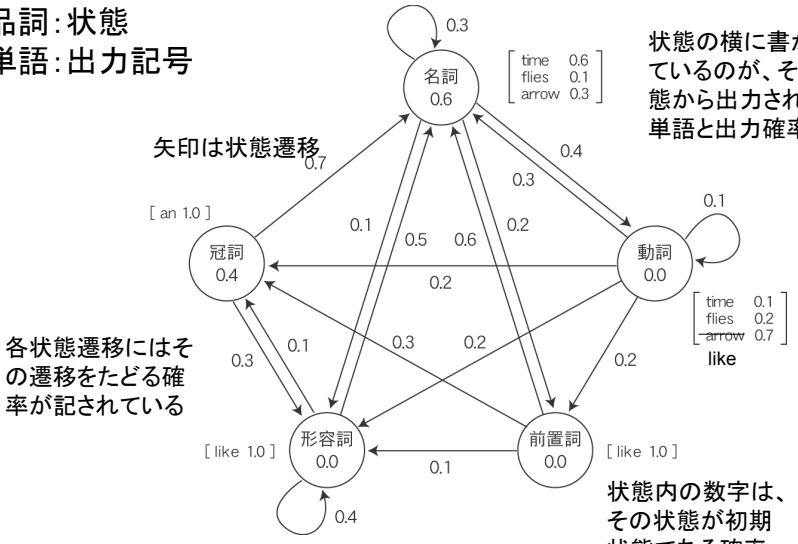
図 5.8　単語とランク

英語の品詞判別の例

- 英文”Time flies like an arrow”の解釈
  - 「光陰矢のごとし」
  - 「時蝿は矢を好む」
  - このようなことが起こるのは、各単語が複数の品詞および意味を持っているため
  - 文を解析する場合には、正しい品詞としてどれを選ぶかという処理が重要になる
- 品詞（名詞・動詞など）を内部状態と捉え、単語を外部から観測できる記号出力と考えると、言語の生成過程は隠れマルコフモデルで近似できる！

単語／品詞の隠れマルコフモデル

·品詞：状態
·単語：出力記号

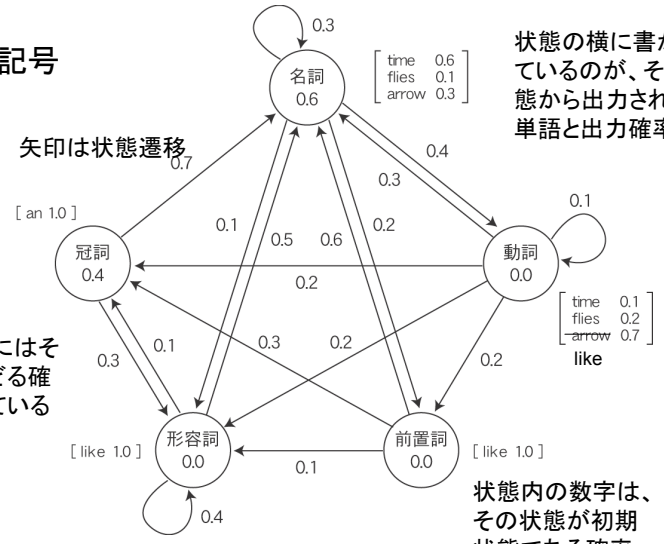


確率

- P_i（名詞）
  - 名詞が初期状態である確率
- P_0（time | 名詞）
  - 名詞の状態から“time”が出力される確率
- P_t（動詞 | 名詞）など
  - 名詞の状態から動詞の状態への遷移確率

単語／品詞の隠れマルコフモデル

・品詞：状態
・単語：出力記号



「光陰矢のごとし」が得られる確率

- 「光陰矢のごとし」
- 「time／名詞, flies／動詞, like／前置詞, an／冠詞, arrow／名詞」が得られる確率P_1

$$
\begin{aligned}
P_1 = {} & P_i(\text{名詞})\, P_0(\text{time} \mid \text{名詞}) \\
& P_t(\text{動詞} \mid \text{名詞})\, P_0(\text{flies} \mid \text{動詞}) \\
& P_t(\text{前置詞} \mid \text{動詞})\, P_0(\text{like} \mid \text{前置詞}) \\
& P_t(\text{冠詞} \mid \text{前置詞})\, P_0(\text{an} \mid \text{冠詞}) \\
& P_t(\text{名詞} \mid \text{冠詞})\, P_0(\text{arrow} \mid \text{名詞}) \\
= {} & \\
= {} &
\end{aligned}
$$

「時蝿は矢を好む」が得られる確率

- 「時蝿は矢を好む」
- 「time／名詞, flies／名詞, like／動詞, an／冠詞, arrow／名詞」が得られる確率P_2

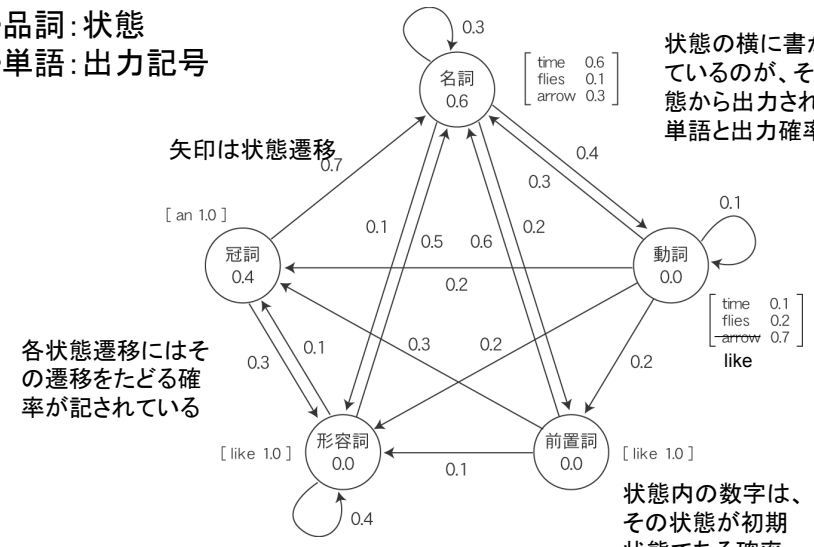
P_2 ＝

＝

＝

単語／品詞の隠れマルコフモデル

·品詞：状態
·単語：出力記号

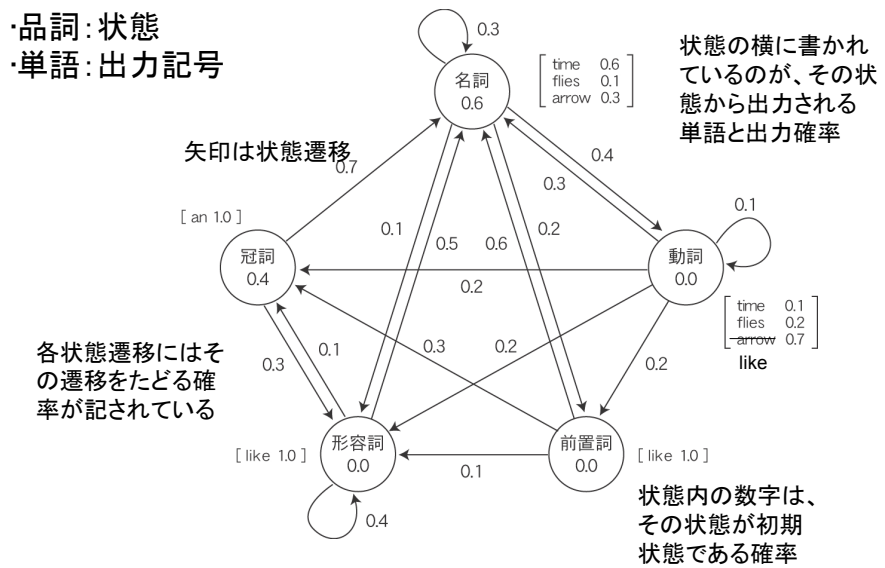


「光陰矢のごとし」が得られる確率

- 「光陰矢のごとし」
- 「time／名詞, flies／動詞, like／前置詞, an／冠詞, arrow／名詞」が得られる確率P_1

$$
\begin{aligned}
P_1 = {} & P_i(\text{名詞})\, P_0(\text{time} \mid \text{名詞}) \\
& P_t(\text{動詞} \mid \text{名詞})\, P_0(\text{flies} \mid \text{動詞}) \\
& P_t(\text{前置詞} \mid \text{動詞})\, P_0(\text{like} \mid \text{前置詞}) \\
& P_t(\text{冠詞} \mid \text{前置詞})\, P_0(\text{an} \mid \text{冠詞}) \\
& P_t(\text{名詞} \mid \text{冠詞})\, P_0(\text{arrow} \mid \text{名詞}) \\
= {} & 0.6 \times 0.6 \times 0.4 \times 0.2 \times 0.2 \times 1.0 \\
& \times 0.3 \times 1.0 \times 0.7 \times 0.3 \\
= {} & 0.0003628
\end{aligned}
$$

単語／品詞の隠れマルコフモデル



「時蝿は矢を好む」が得られる確率

- 「時蝿は矢を好む」
- 「time／名詞, flies／名詞, like／動詞, an／冠詞, arrow／名詞」が得られる確率P_2

$P_2 = P_i$（名詞）P_0（time｜名詞）
P_t（名詞｜名詞）P_0（flies｜名詞）
P_t（動詞｜名詞）P_0（like｜動詞）
P_t（冠詞｜動詞）P_0（an｜冠詞）
P_t（名詞｜冠詞）P_0（arrow｜名詞）
$= 0.6 \times 0.6 \times 0.3 \times 0.1 \times 0.4 \times 0.7 \times 0.2 \times 1.0 \times 0.7 \times 0.3$
$= 0.0001270$

このモデルからわかること

- さきほどの隠れマルコフモデルからは、全部で6個の状態遷移系列（単語／品詞系列）が生成される
- 文“Time flies like an arrow”の生成確率は、これら6個の系列の確率の和として求めることができる。
- また、これらの系列の中で最も高い確率を与えるものが、この文に対する最適な品詞付けであると解釈することができる。

隠れマルコフモデル
の形式と特徴

マルコフモデル

- ある記号の出現確率が直前の記号のみに依存すると仮定する確率モデルを「マルコフモデル」という。

- 確率変数の系列 X1, X2, ・・・を考え、これらの確率変数のとりうる値の集合をQ={q1, ・・・, qn} とする。
- 確率変数Xnの実現値xn (xn ∈Q)を時系列nにおける「状態」と呼び、Xn=xnで表す
- このような系列{Xn}は確率過程（stochastic process）と呼ばれる。
- 確率過程を特徴付けるためには、一般に次の2つを知る必要がある。

① P(X1=q1)：各状態qiの初期状態確率
② P(Xn=xn｜X1=x1,・・・,Xn-1=xn-1)：過去の状態系列に対する次の状態の条件付き確率

隠れマルコフモデル

- マルコフ過程の各状態において、確率的な記号の出力を考えたモデル
- 5項組M=(Q, Σ ,A,B, π)により定義される。

① Q={q1,・・・,qN}：状態の有限集合
② Σ={o1,・・・,oM}：出力記号の有限集合
③ A={aij}：状態遷移確率分布
（aijは、状態qiから状態qjへの遷移確率であり、Σj aij=1を満たす）
④ B={bi(ot)}：記号出力確率分布
（bi(ot)は、状態qiで記号otを出力する確率であり、Σt bi(ot)=1を満たす）
⑤ π={πi}：初期状態確率分布
（πiは、状態qiが初期状態である確率P(X1=qi)である。

二重の確率モデル

- 隠れマルコフモデルは、状態間の遷移が確率的であり、また各状態における出力もある確率分布に従って現れるという点で、2重に確率的なモデルである。
- 隠れマルコフモデルは、一方で、確率モデルとしてはベイジアンネットワークの特殊形といえる。
- 他方では有限オートマトンの状態遷移が確率的に起こるとした拡張版と考えることもできる。

隠れマルコフモデルの例②
日本語の品詞判別

かな漢字変換

- ひらがな表記された文字列を漢字かな混じり表記に変換する

- 「へんなじがでる」をかな漢字変換すると・・・
  - へんな/形容詞 じが/名詞 でる/動詞
    → 変な自我出る, 変な自画出る ,..
  - へんな/形容詞 じ/名詞 が/助詞 でる/動詞
    → 変な字が出る, 変な痔が出る, ...

かな漢字変換

- かな漢字変換プログラムの仕事
  (1)ひらがなで表記された入力文を単語に分割すること
  (2)それぞれの単語の可能な漢字表記（同音異義語）の中で最も妥当なものを選ぶということ

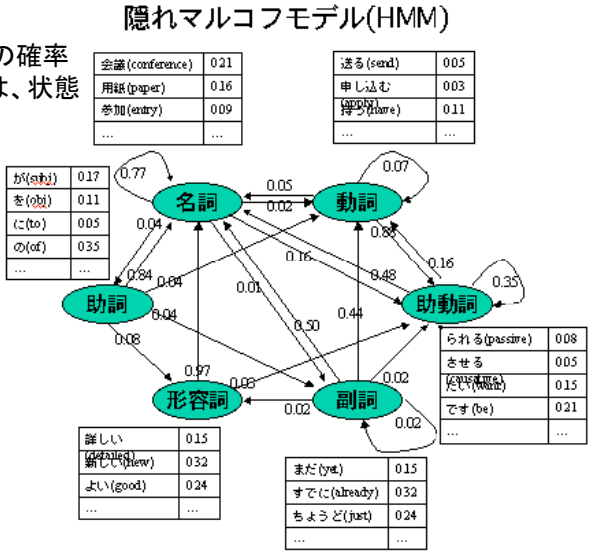
- 失敗例
  (1)に失敗すると「変な自我出る」になり
  (2)に失敗すると「変な痔が出る」になる。

かな漢字変換

- かな漢字変換で正解を得る鍵は、
- 複数の解釈の可能性の中から日本語として最も妥当な解釈を選択するための判断基準、
- すなわち、日本語の「文法」をコンピュータ上で表現する方法にある。

日本語の「文法」を表現した隠れマルコフモデル

·ノードは内部状態（品詞）
·リンクは状態遷移およびその確率
·ノードに付属するテーブルは、状態別の記号(単語)の出現確率



かな漢字変換

- 標準的な日本語では、名詞の直後には動詞より助詞が接続する可能性が高い。隠れマルコフモデルでは、このような単語の接続の自然性を状態遷移確率の大小で表現する。
  - 「へんなじがでる」のかな漢字変換において、「変な自我出る」を不自然と感じる主な原因は、「自我」という名詞の直後に「出る」という動詞が接続し、助詞が省略されているせいである。
- 「じ」の変換候補としては「痔」より「字」の方が可能性が高いことは、名詞という内部状態における出現確率の大小で表現する。

参考文献

- 『確率的言語モデル』(北 研二, 東京大学出版会, 1999)
  - 第4章 隠れマルコフモデル
- 「ゲノムや言語をどう読み解くか？　言語とゲノムの意外な関係―かな漢字変換と遺伝子発見 」(永田昌明,生命誌ジャーナル 2002年号)
  - http://www.brh.co.jp/experience/exhibition/journal/33/resarch_21.html

Keio University SFC 2004

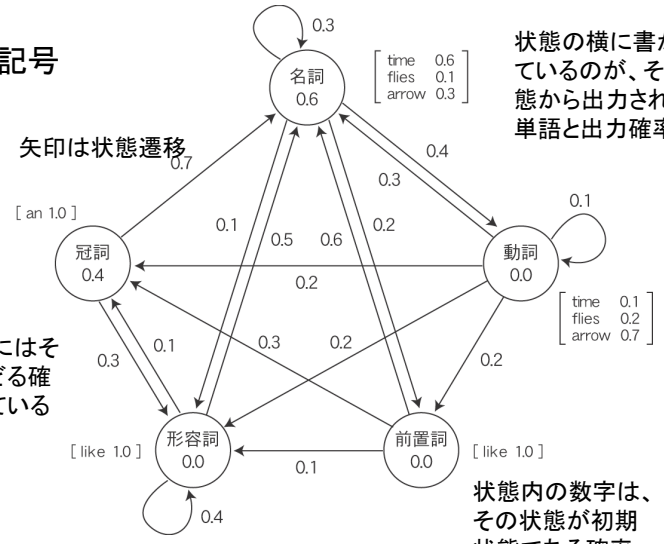
『探索的モデリング』

第11回 隠れマルコフモデル①

いば　たかし
井庭 崇
慶應義塾大学総合政策学部 専任講師
iba@sfc.keio.ac.jp
http://www.sfc.keio.ac.jp/~iba/lecture/

単語／品詞の隠れマルコフモデル

·品詞：状態
·単語：出力記号



「光陰矢のごとし」が得られる確率

- 「光陰矢のごとし」
- 「time／名詞, flies／動詞, like／前置詞, an／冠詞, arrow／名詞」が得られる確率P_1

$$
\begin{aligned}
P_1 = {} & P_i(\text{名詞})\, P_0(\text{time} \mid \text{名詞}) \\
& P_t(\text{動詞} \mid \text{名詞})\, P_0(\text{flies} \mid \text{動詞}) \\
& P_t(\text{前置詞} \mid \text{動詞})\, P_0(\text{like} \mid \text{前置詞}) \\
& P_t(\text{冠詞} \mid \text{前置詞})\, P_0(\text{an} \mid \text{冠詞}) \\
& P_t(\text{名詞} \mid \text{冠詞})\, P_0(\text{arrow} \mid \text{名詞}) \\
= {} & \\
= {} &
\end{aligned}
$$

「時蝿は矢を好む」が得られる確率

- 「時蝿は矢を好む」
- 「time／名詞, flies／名詞, like／動詞, an／冠詞, arrow／名詞」が得られる確率P_1

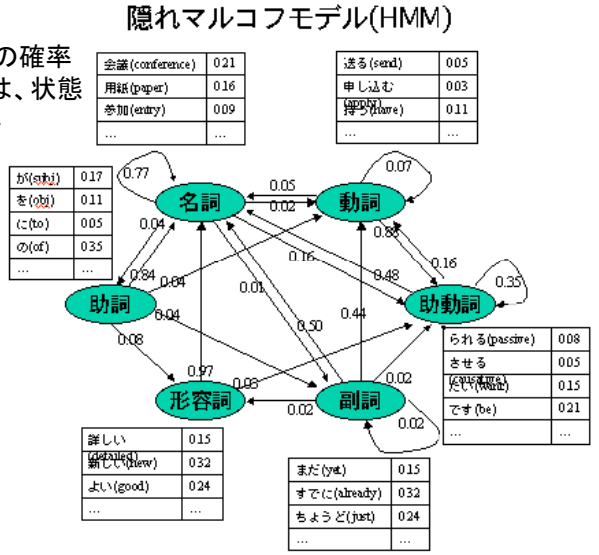
P_1 ＝

＝

＝

日本語の「文法」を表現した隠れマルコフモデル

·ノードは内部状態（品詞）
·リンクは状態遷移およびその確率
·ノードに付属するテーブルは、状態別の記号(単語)の出現確率



「光陰矢のごとし」が得られる確率

- 「光陰矢のごとし」
- 「time／名詞, flies／動詞, like／前置詞, an／冠詞, arrow／名詞」が得られる確率P_1

$P_1 = P_i(\text{名詞})\, P_0(\text{time} \mid \text{名詞})$
$P_t(\text{動詞} \mid \text{名詞})\, P_0(\text{flies} \mid \text{動詞})$
$P_t(\text{前置詞} \mid \text{動詞})\, P_0(\text{like} \mid \text{前置詞})$
$P_t(\text{冠詞} \mid \text{前置詞})\, P_0(\text{an} \mid \text{冠詞})$
$P_t(\text{名詞} \mid \text{冠詞})\, P_0(\text{arrow} \mid \text{名詞})$
$= 0.6 \times 0.6 \times 0.4 \times 0.2 \times 0.2 \times 1.0 \times 0.3 \times 1.0 \times 0.7 \times 0.3$
$= 0.0003628$

「時蝿は矢を好む」が得られる確率

- 「時蝿は矢を好む」
- 「time／名詞, flies／名詞, like／動詞, an／冠詞, arrow／名詞」が得られる確率P_2

$P_2 = P_i(\text{名詞})\, P_0(\text{time} \mid \text{名詞})$
$P_t(\text{名詞} \mid \text{名詞})\, P_0(\text{flies} \mid \text{名詞})$
$P_t(\text{動詞} \mid \text{名詞})\, P_0(\text{like} \mid \text{動詞})$
$P_t(\text{冠詞} \mid \text{動詞})\, P_0(\text{an} \mid \text{冠詞})$
$P_t(\text{名詞} \mid \text{冠詞})\, P_0(\text{arrow} \mid \text{名詞})$
$= 0.6 \times 0.6 \times 0.3 \times 0.1 \times 0.4 \times 0.7 \times 0.2 \times 1.0 \times 0.7 \times 0.3$
$= 0.0001270$