# Скрытые марковские модели

Разрешение морфологической неоднозначности

# Постморфологический анализ

- =предсинтаксический анализ
- Предназначен для устранения морфологической омонимии (многозначности) слов
  –Выбор правильной леммы
  –Уточнение морфологических характеристик

**Основные методы**

- Написание правил,
- Статистические методы, прежде всего, на основе морфологически размеченного корпуса

**-Скрытые марковские модели**

# Марковские модели

- Набор состояний: $\{s_1, s_2, \ldots, s_N\}$
- Процесс движется от одного состояния к другому, порождая последовательность состояний :
- $s_{i1}, s_{i2}, \ldots, s_{ik}, \ldots$
- Свойство марковской цепи: вероятность следующего состояния зависит от состояния предыдущего:

$$P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1})$$

- Чтобы определить марковскую сеть, должны быть определены следующие вероятности

$$\pi_i = P(s_i) \qquad a_{ij} = P(s_i \mid s_j)$$

# Пример марковской модели


- Два состояния : ‘Rain’ и ‘Dry’
- Вероятности переходов: P(‘Rain’|‘Rain’)=0.3 , P(‘Dry’|‘Rain’)=0.7 , P(‘Rain’|‘Dry’)=0.2, P(‘Dry’|‘Dry’)=0.8
- Исходные вероятности:  P(‘Rain’)=0.4 , P(‘Dry’)=0.6

# Вычисление вероятности последовательности

- По свойству марковской цепи, вероятность последовательности состояний может быть найдена по формуле

$$P(s_{i1}, s_{i2}, \ldots, s_{ik}) = P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) P(s_{i1}, s_{i2}, \ldots, s_{ik-1})$$
$$= P(s_{ik} \mid s_{ik-1}) P(s_{i1}, s_{i2}, \ldots, s_{ik-1}) = \ldots$$
$$= P(s_{ik} \mid s_{ik-1}) P(s_{ik-1} \mid s_{ik-2}) \ldots P(s_{i2} \mid s_{i1}) P(s_{i1})$$

- Предположим, мы хотим подсчитать вероятность последовательности: {‘Dry’,’Dry’,’Rain’,Rain’}.

$$\mathrm{P}(\{\text{‘Dry’,’Dry’,’Rain’,Rain’}\}) =$$
$$\mathrm{P}(\text{‘Rain’}|\text{’Rain’})\ \mathrm{P}(\text{‘Rain’}|\text{’Dry’})\ \mathrm{P}(\text{‘Dry’}|\text{’Dry’})\ \mathrm{P}(\text{‘Dry’}) =$$
$$= 0.3*0.2*0.8*0.6$$

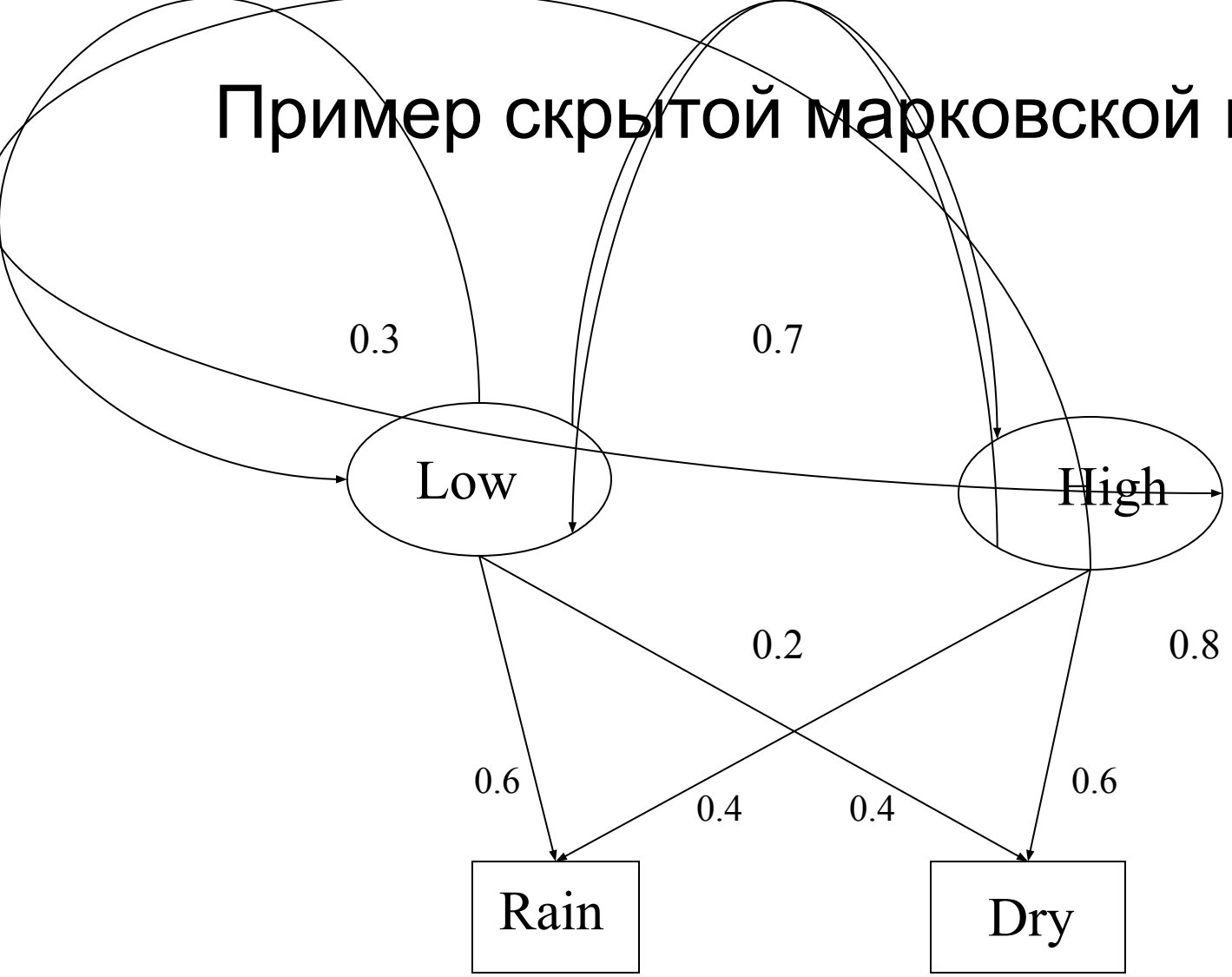
# Скрытые марковские модели

- Множество состояний: $\{s_1, s_2, \ldots, s_N\}$
- Процесс движется от состояния к состоянию :
- Выполняется свойство марковской цепи:
- Состояния – невидимы, но каждое состояние порождает одно из M наблюдений - видимых состояний

$$\{v_1, v_2, \ldots, v_M\}$$

- Чтобы определить скрытую марковскую цепь, нужно определить
  - Матрицу переходов $A=(a_{ij})$, $a_{ij}= P(s_i \mid s_j)$ ,
  - Матрицу вероятностей наблюдаемых состояний $B=(b_i(v_m))$, $b_i(v_m) = P(v_m \mid s_i)$
  - Вектор начальных вероятностей $\pi=(\pi_i)$, $\pi_i = P(s_i)$.
  - Модель представлена $M=(A, B, \pi)$.

# Пример скрытой марковской модели



# Пример скрытой марковской модели

- Два состояния: ‘Низкое’ and ‘Высокое’ атм. давление.
- Два наблюдения: ‘Дождь’ and ‘Сухо’.
- Вероятности перехода: $P(\text{‘Low’}|\text{‘Low’})=0.3$ ,
  $P(\text{‘High’}|\text{‘Low’})=0.7$ , $P(\text{‘Low’}|\text{‘High’})=0.2$,
  $P(\text{‘High’}|\text{‘High’})=0.8$
- Вероятности наблюдения: $P(\text{‘Rain’}|\text{‘Low’})=0.6$ ,
  $P(\text{‘Dry’}|\text{‘Low’})=0.4$ , $P(\text{‘Rain’}|\text{‘High’})=0.4$ ,
  $P(\text{‘Dry’}|\text{‘High’})=0.3$ .
- Начальные вероятности: $P(\text{‘Low’})=0.4$ , $P(\text{‘High’})=0.6$ .

# Пример вычисления вероятности наблюдений

- Хотим вычислить вероятность последовательности, {‘Dry’,’Rain’}.
- Рассмотрим все возможные скрытые состояния:

$$P(\{\text{‘Dry’,’Rain’}\}) = P(\{\text{‘Dry’,’Rain’}\}, \{\text{‘Low’,’Low’}\}) + P(\{\text{‘Dry’,’Rain’}\}, \{\text{‘Low’,’High’}\}) + P(\{\text{‘Dry’,’Rain’}\}, \{\text{‘High’,’Low’}\}) + P(\{\text{‘Dry’,’Rain’}\}, \{\text{‘High’,’High’}\})$$

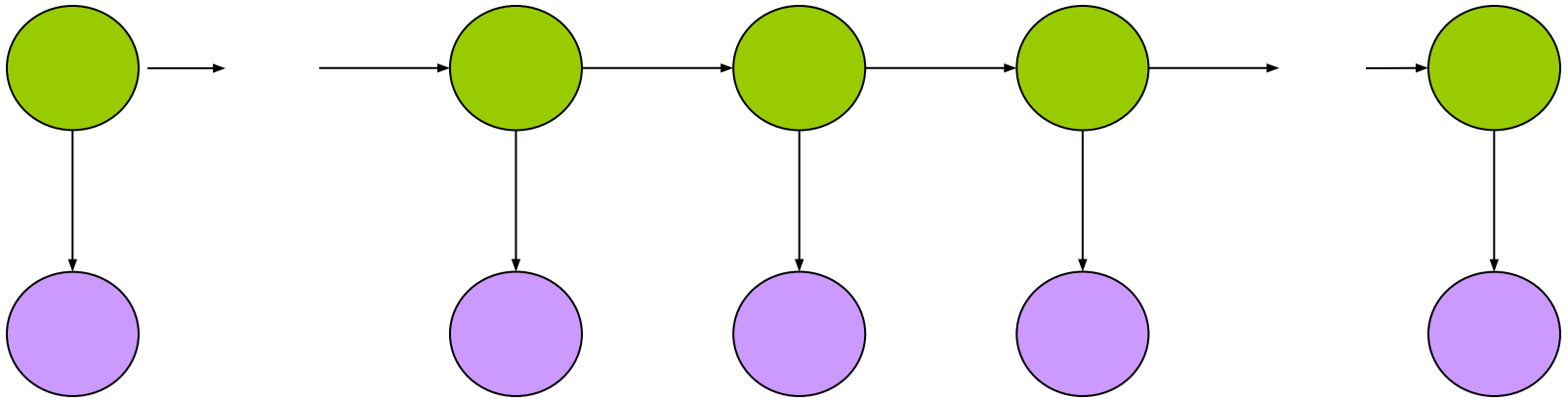
где первый элемент:

$$P(\{\text{‘Dry’,’Rain’}\}, \{\text{‘Low’,’Low’}\}) = P(\{\text{‘Dry’,’Rain’}\} \mid \{\text{‘Low’,’Low’}\})\ P(\{\text{‘Low’,’Low’}\}) = P(\text{‘Dry’}|\text{’Low’})P(\text{‘Rain’}|\text{’Low’})\ P(\text{‘Low’})P(\text{‘Low’}|\text{’Low}) = 0.4*0.6*0.4*0.3$$

# Почему важно рассмотрение НММ в автоматической обработке текста

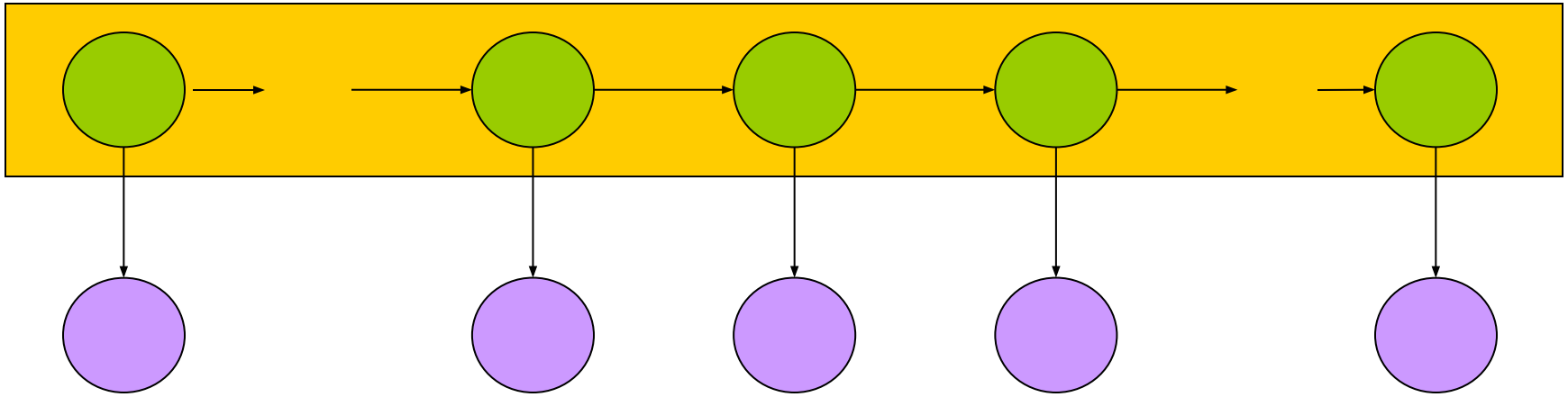
- Непосредственно имеем дело с неоднозначными словами и конструкциями
- Нужно распознавать скрытые
  - Части речи
  - Лексические значения
  - Типы именованных сущностей (организация, персона, географическое место …)
  - Определение тональности предложения
  - и др.

# Что такое HMM?



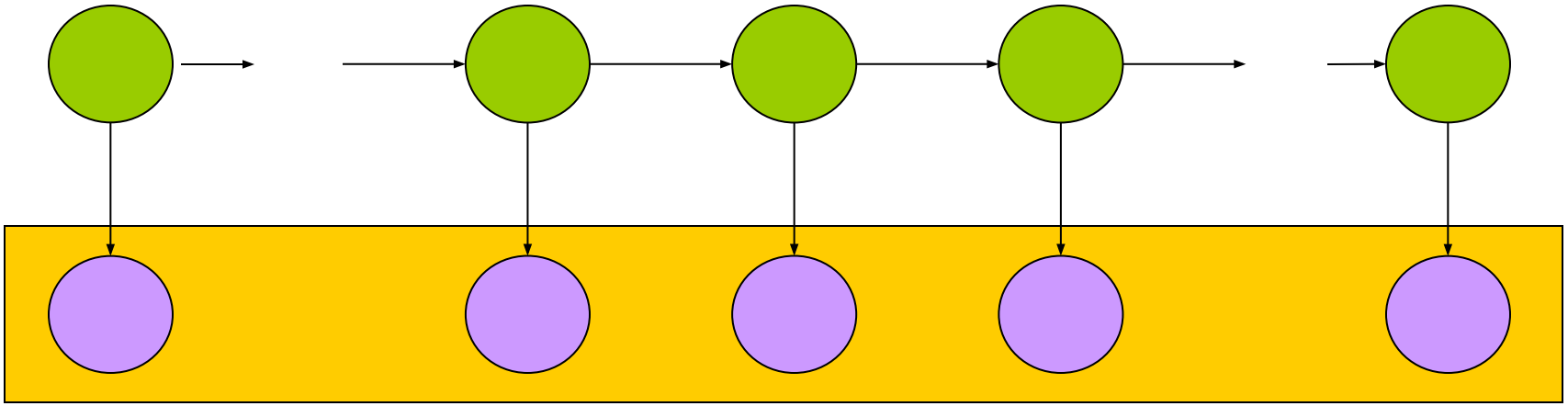
- Графическая модель
- Кружки – это состояния
- Стрелки обозначают вероятностные зависимости между состояниями

# Что такое HMM?



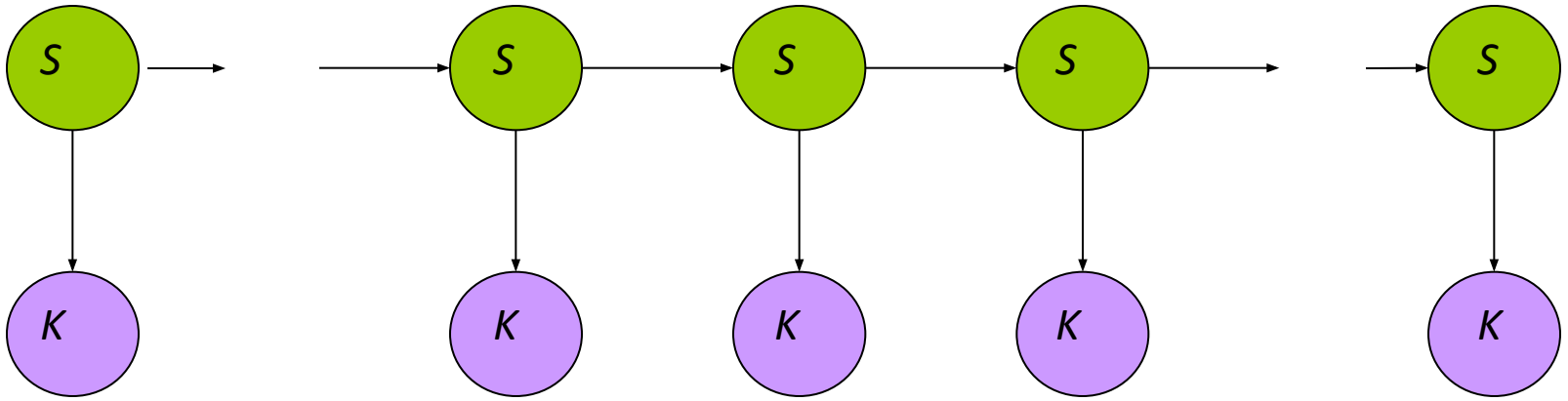
- Зеленые кружки – это скрытые состояния
- Зависят только от предыдущего состояния

# Что такое HMM?



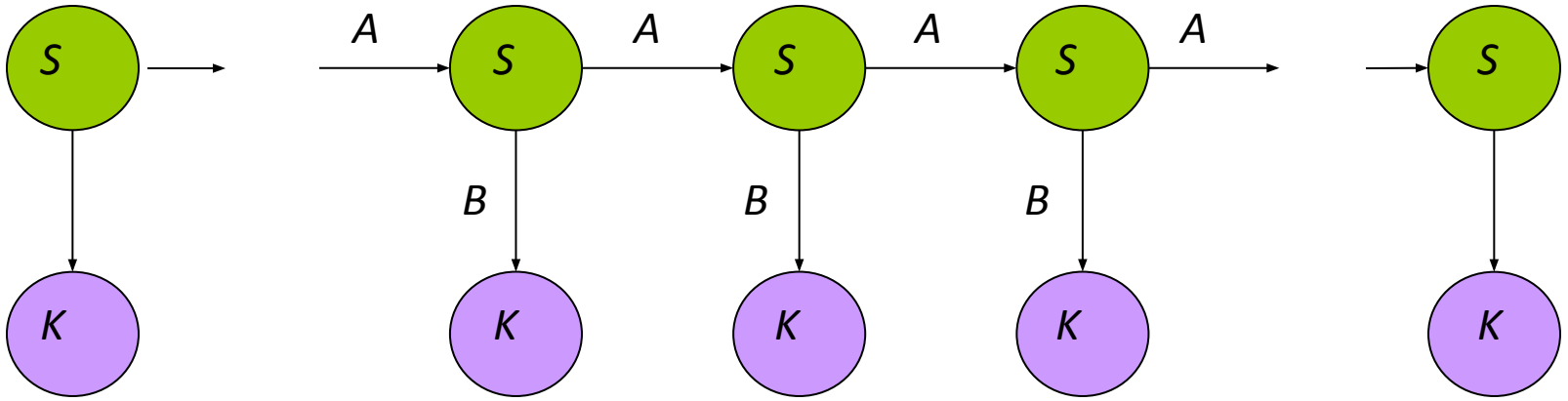
- Фиолетовые кружки – это наблюдаемые состояния
- Зависят только от соответствующих скрытых состояний

# HMM формализм



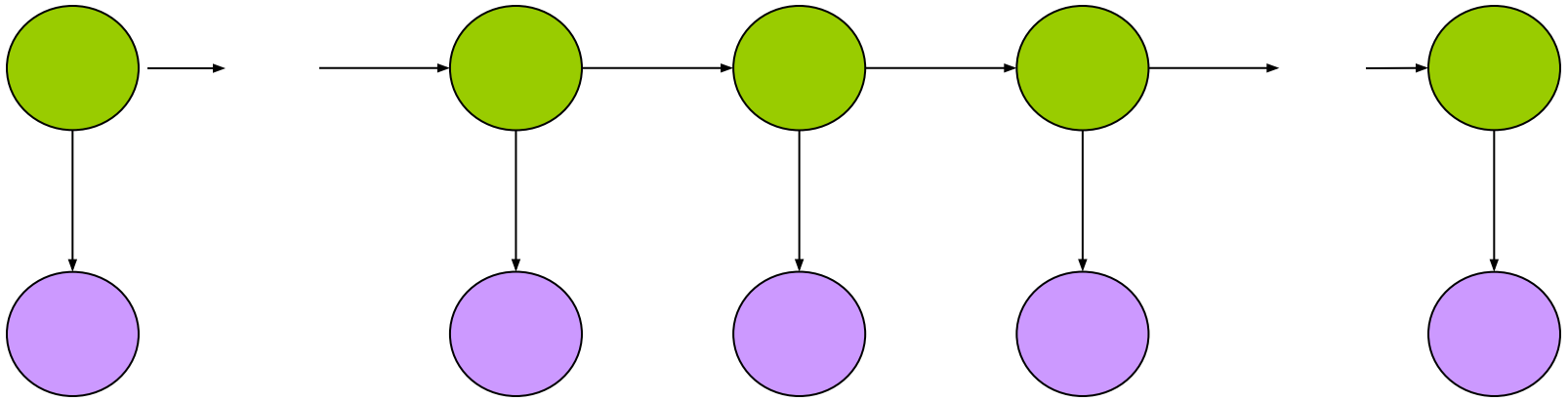

- $\{S, K, \Pi, A, B\}$
- $S : \{s_1 \ldots s_N\}$ - значения скрытых состояний
- $K : \{k_1 \ldots k_M\}$ – значения наблюдаемых состояний

# HMM формализм



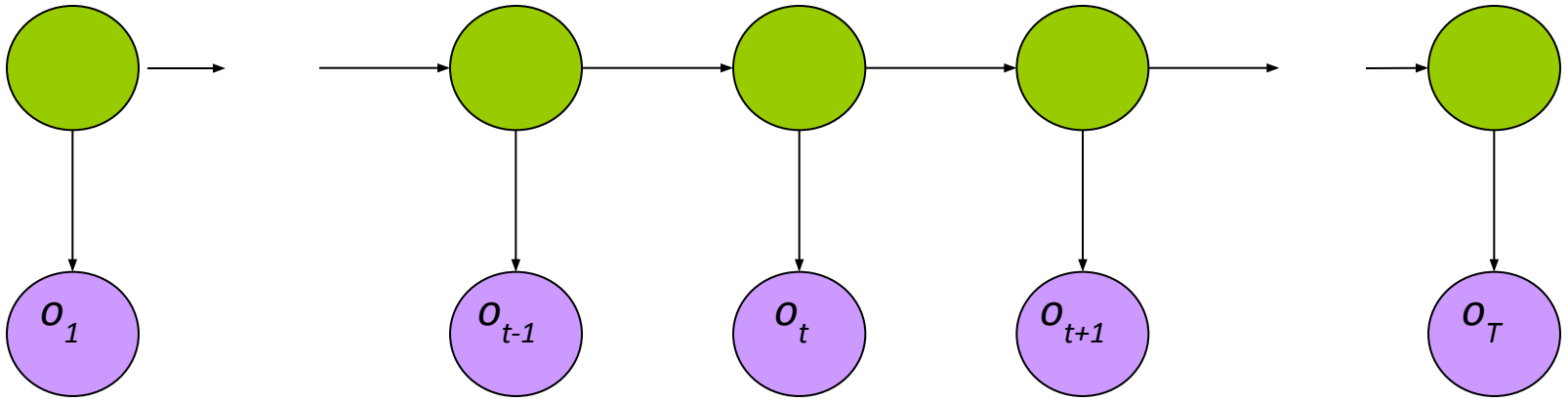

- $\{S, K, \Pi, A, B\}$
- $\Pi = \{\pi_i\}$ - вероятности начальных состояний
- $A = \{a_{ij}\}$ - вероятности переходов между скрытыми состояниями
- $B = \{b_{ik}\}$ – вероятности наблюдаемых состояний

# Вывод HMM



- Вычислить вероятность последовательности наблюдаемых состояний (**Evaluation**)
- Имея последовательность наблюдаемых состояний, вычислить наиболее вероятную последовательность скрытых состояний (**Decoding**)
- Имея последовательность наблюдаемых состояний и множество возможных моделей, определить какая модель лучше соответствует данным (т.е. наблюдаемой последовательности) (**Learning**)

# Оценка (Evaluation)

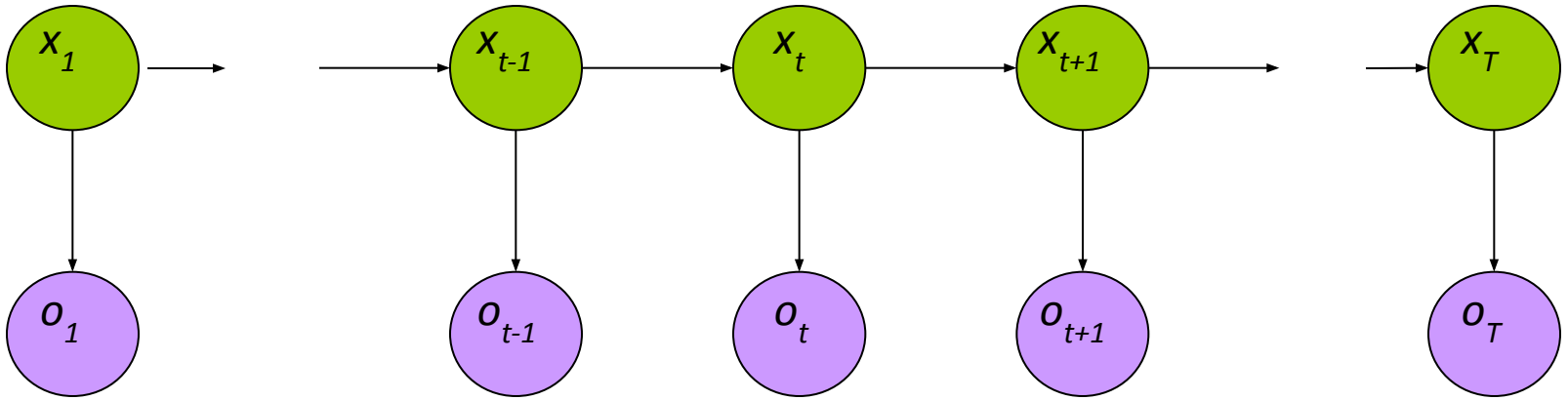



Имея последовательность наблюдаемых состояний и модель, вычислить вероятность последовательности наблюдаемых состояний

$$O = (o_1 \ldots o_T),\ \mu = (A, B, \Pi)$$

Вычислить $P(O \mid \mu)$

# Оценка (Evaluation)




$$P(O \mid \mu) = \sum_{\{x_1 \ldots x_T\}} \pi_{x_1} b_{x_1 o_1} \prod_{t=1}^{T-1} a_{x_t x_{t+1}} b_{x_{t+1} o_{t+1}}$$

Сложность O ($N^T$), где N – число возможных вариантов состояний

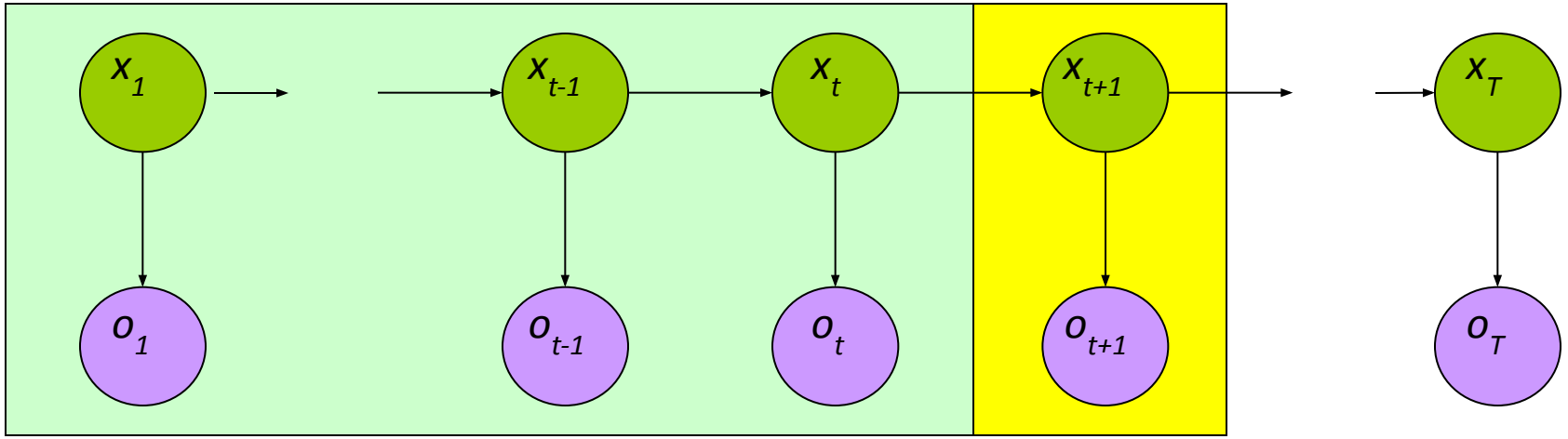
# Форвардная процедура

- Метод динамического программирования
- Определим переменную:

$$\alpha_i(t) = P(o_1 \ldots o_t, x_t = i \mid \mu)$$

Смысл переменной α: вероятность наблюдений $o_1, \ldots o_t$ и при этом оказаться в состоянии i

# Форвардная процедура




$$\alpha_j(t+1)$$

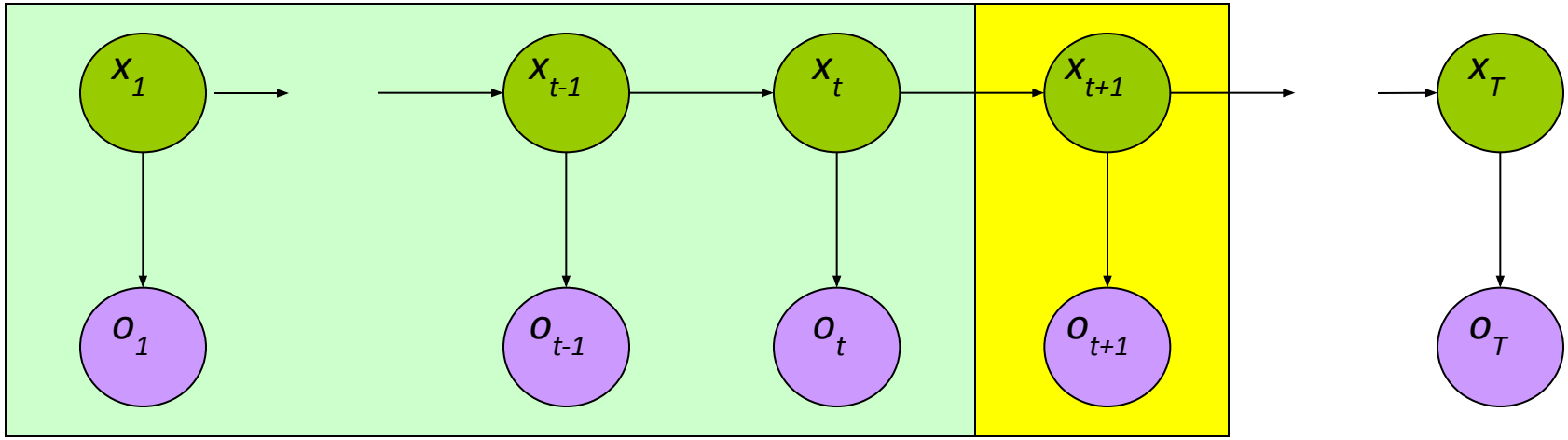
$$= P(o_1 \ldots o_{t+1}, x_{t+1} = j)$$

$$= P(o_1 \ldots o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t \mid x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j) P(x_{t+1} = j)$$

$$= P(o_1 \ldots o_t, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

# Форвардная процедура




$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i, x_{t+1} = j) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_{t+1} = j \mid x_t = i) P(x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t, x_t = i) P(x_{t+1} = j \mid x_t = i) P(o_{t+1} \mid x_{t+1} = j)$$

$$= \sum_{i=1\ldots N} \alpha_i(t) a_{ij} b_{jo_{t+1}}$$

# Вычисление вероятности последовательности наблюдаемых событий

- Можем эффективно вычислять
- $\alpha_T(I)=P(o_1, o_2,\ldots o_T, x_T=i|\mu)$

- Как вычислить
- $P(o_1, o_2,\ldots o_T |\mu)$?

- Как вычислить
- $P(x_T=i|o_1, o_2,\ldots o_T ,\mu)$?

# Вычисление вероятности последовательности наблюдаемых событий

- Можем эффективно вычислять
- $\alpha_T(i)=P(o_1, o_2,\ldots o_T, x_T=i|\mu)$

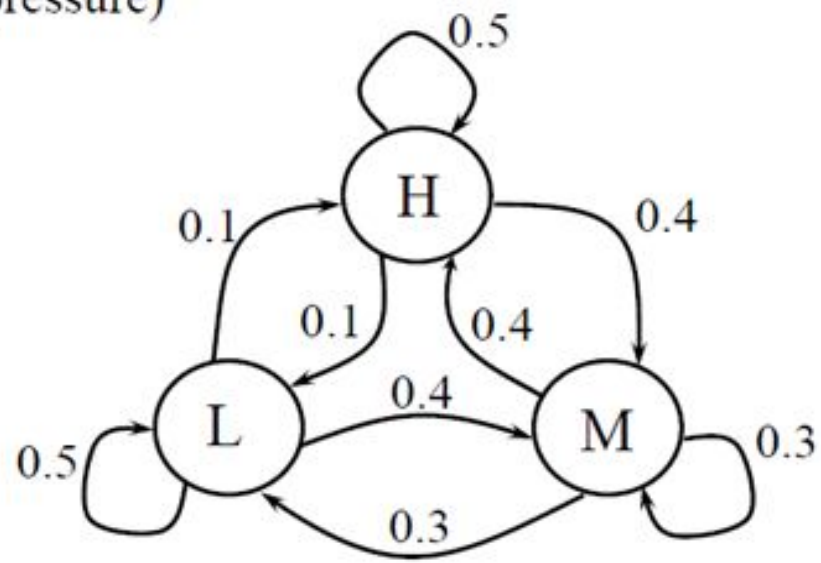
- Как вычислить
- $P(o_1, o_2,\ldots o_T |\mu) = \Sigma_i \alpha_T(i)$

- Как вычислить
- $P(x_T=i|o_1, o_2,\ldots o_T ,\mu)= \alpha_T(i)/(\Sigma_i \alpha_T(i))$

# Форвардный алгоритм: пример

- Given the following model of the weather:

(states indicate pressure)




| | state M | state H | state L |
|---|---|---|---|
| $P$(sun) | 0.50 | 0.75 | 0.25 |
| $P$(rain) | 0.50 | 0.25 | 0.75 |

$\pi_M = 0.50$
$\pi_H = 0.20$
$\pi_L = 0.30$

# Форвардный алгоритм

- Найти вероятность последовательности:
- s r r s r (s- sun, r – rain)

## Forward Algorithm: An Example

- What is the probability of the observation sequence:
  s r r s r (s=sun,r=rain)?

- Compute value of $\Sigma\alpha$ at time 5...

$\alpha_1(M)=0.5\cdot0.5$ $\alpha_1(H)=0.2\cdot0.75$ $\alpha_1(L)=0.3\cdot0.25$
$\alpha_1(M)=0.25$ $\alpha_1(H)=0.15$ $\alpha_1(L)=0.075$

$$\alpha_2(M) = [0.25\cdot0.3 + 0.15\cdot0.4 + 0.075\cdot0.4] \cdot 0.5 = 0.0825$$
$$\alpha_2(H) = [0.25\cdot0.4 + 0.15\cdot0.5 + 0.075\cdot0.1] \cdot 0.25 = 0.0456$$
$$\alpha_2(L) = [0.25\cdot0.3 + 0.15\cdot0.1 + 0.075\cdot0.5] \cdot 0.75 = 0.0956$$

$$\alpha_3(M) = [0.0825\cdot0.3 + 0.0456\cdot0.4 + 0.0956\cdot0.4] \cdot 0.5 = 0.0406$$
$$\alpha_3(H) = [0.0825\cdot0.4 + 0.0456\cdot0.5 + 0.0956\cdot0.1] \cdot 0.25 = 0.0163$$
$$\alpha_3(L) = [0.0825\cdot0.3 + 0.0456\cdot0.1 + 0.0956\cdot0.5] \cdot 0.75 = 0.0578$$

## Forward Algorithm: An Example

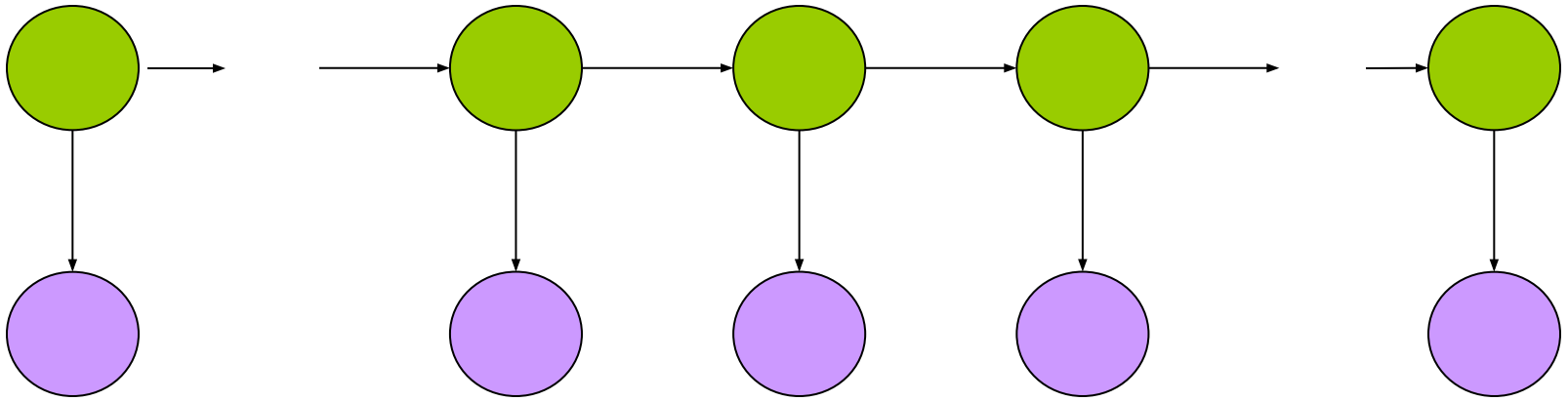
$$\alpha_4(M) = [0.0406 \cdot 0.3 + 0.0163 \cdot 0.4 + 0.0578 \cdot 0.4] \cdot 0.5 = 0.0209$$
$$\alpha_4(H) = [0.0406 \cdot 0.4 + 0.0163 \cdot 0.5 + 0.0578 \cdot 0.1] \cdot 0.75 = 0.0226$$
$$\alpha_4(L) = [0.0406 \cdot 0.3 + 0.0163 \cdot 0.1 + 0.0578 \cdot 0.5] \cdot 0.25 = 0.0107$$

$$\alpha_5(M) = [0.0209 \cdot 0.3 + 0.0226 \cdot 0.4 + 0.0107 \cdot 0.4] \cdot 0.5 = 0.0098$$
$$\alpha_5(H) = [0.0209 \cdot 0.4 + 0.0226 \cdot 0.5 + 0.0107 \cdot 0.1] \cdot 0.25 = 0.0052$$
$$\alpha_5(L) = [0.0209 \cdot 0.3 + 0.0226 \cdot 0.1 + 0.0107 \cdot 0.5] \cdot 0.75 = 0.0104$$

---

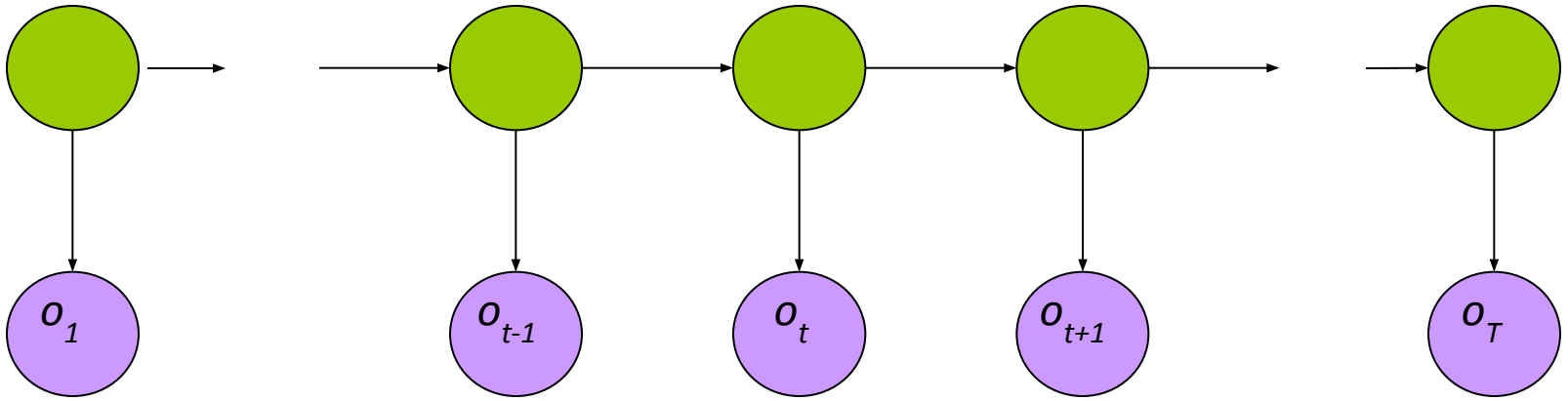
**0.0254**

# Декодирование



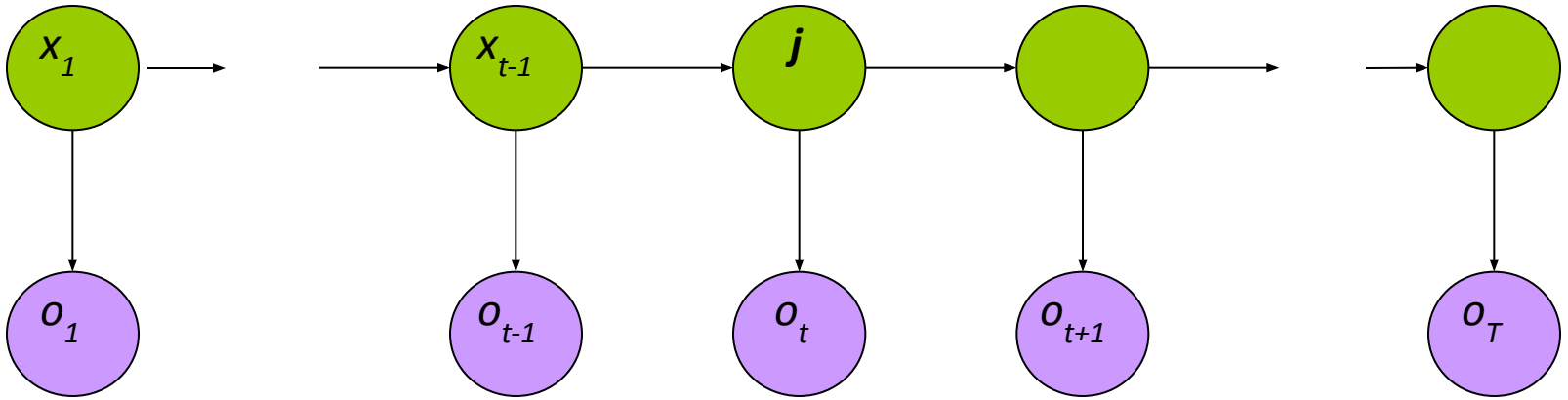
- Вычислить вероятность последовательности наблюдаемых состояний (**Evaluation**)
- Имея последовательность наблюдаемых состояний, вычислить наиболее вероятную последовательность скрытых состояний (**Decoding**)
- Имея последовательность наблюдаемых состояний и множество возможных моделей, определить какая модель лучше соответствует данным (т.е. наблюдаемой последовательности) (**Learning**)

# Декодирование:
## Best State Sequence




- Найти множество состояний, которые наилучшим образом объясняют последовательность видимых состояний
- **Viterbi** algorithm
- $\arg\max_X P(X \mid O)$

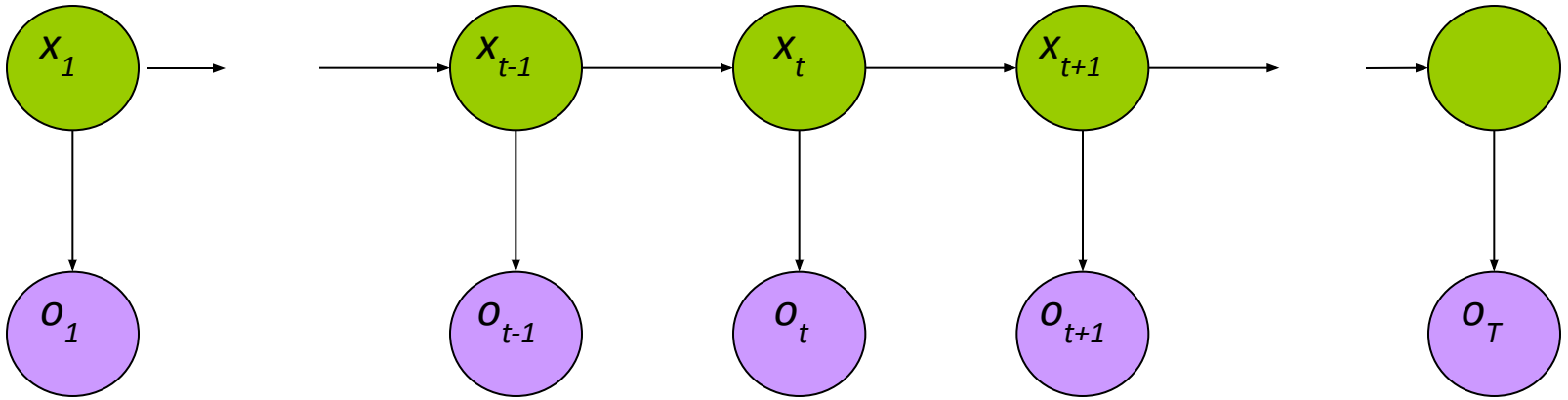
# Алгоритм Витерби




$$\delta_j(t) = \max_{x_1 \ldots x_{t-1}} P(x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, x_t = j, o_t)$$

Последовательность состояний, которая максимизирует вероятность увидеть заданную последовательность видимых состояний во время t-1, остановиться в состоянии j, и увидеть заданное наблюдение во время t

# Алгоритм Витерби



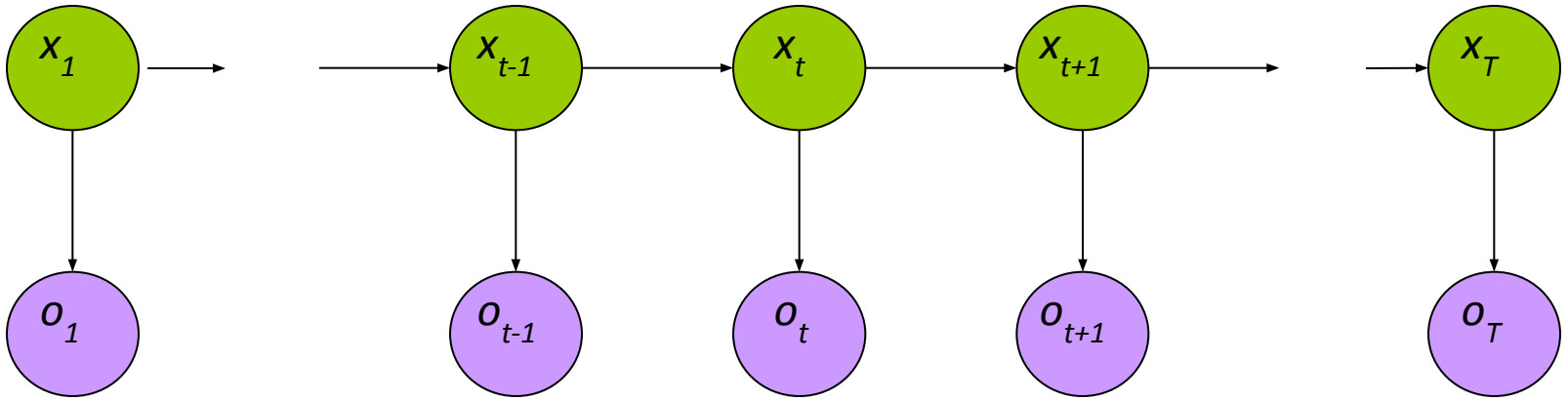

$$\delta_j(t) = \max_{x_1 \ldots x_{t-1}} P(x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, x_t = j, o_t)$$

$$\delta_j(t+1) = \max_i \delta_i(t) a_{ij} b_{jo_{t+1}}$$

$$\psi_j(t+1) = \arg\max_i \delta_i(t) a_{ij} b_{jo_{t+1}}$$

Рекурсивное вычисление

# Алгоритм Витерби



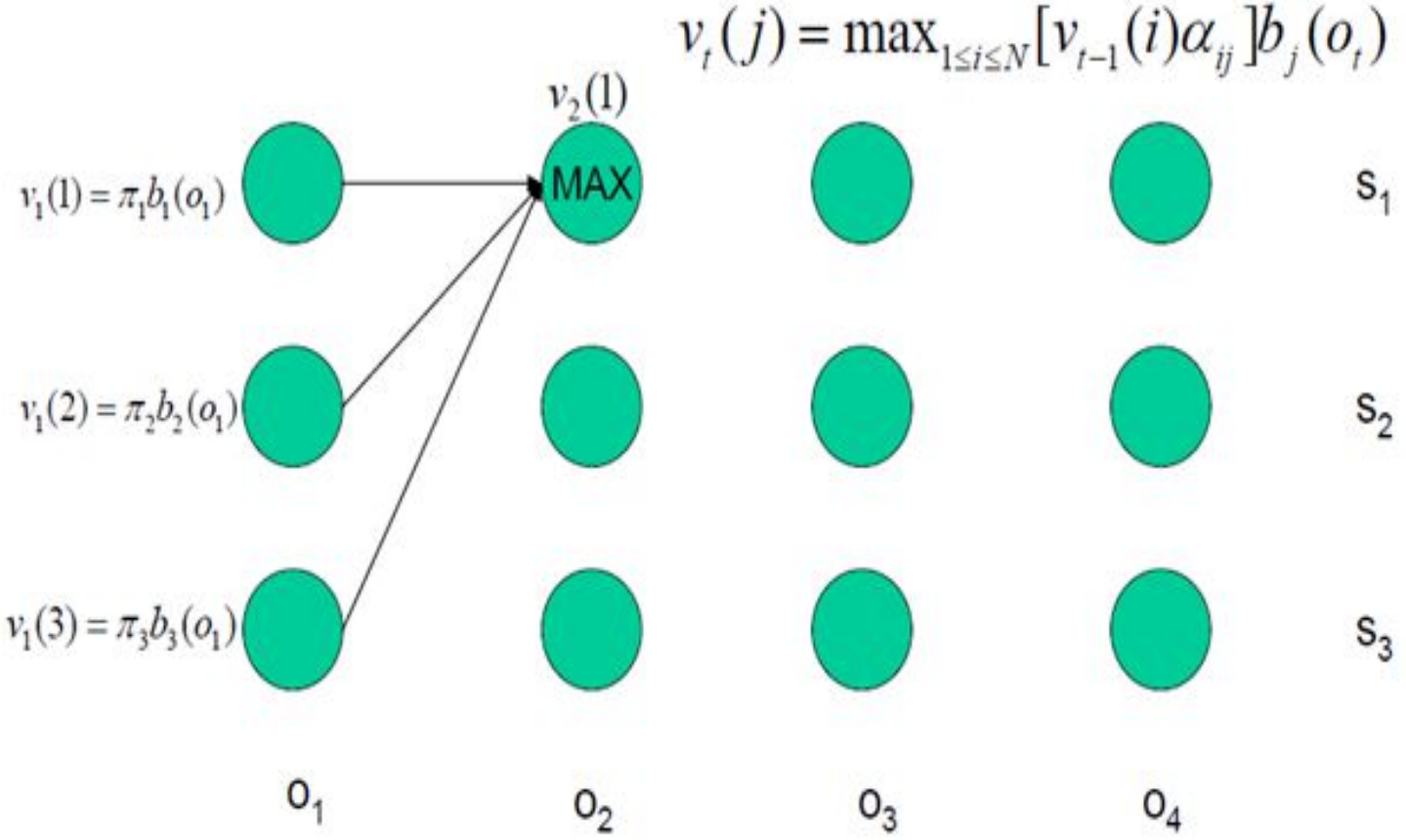

$$\hat{X}_T = \arg\max_i \delta_i(T)$$

$$\hat{X}_t = \psi_{\hat{X}_{t+1}}(t+1)$$

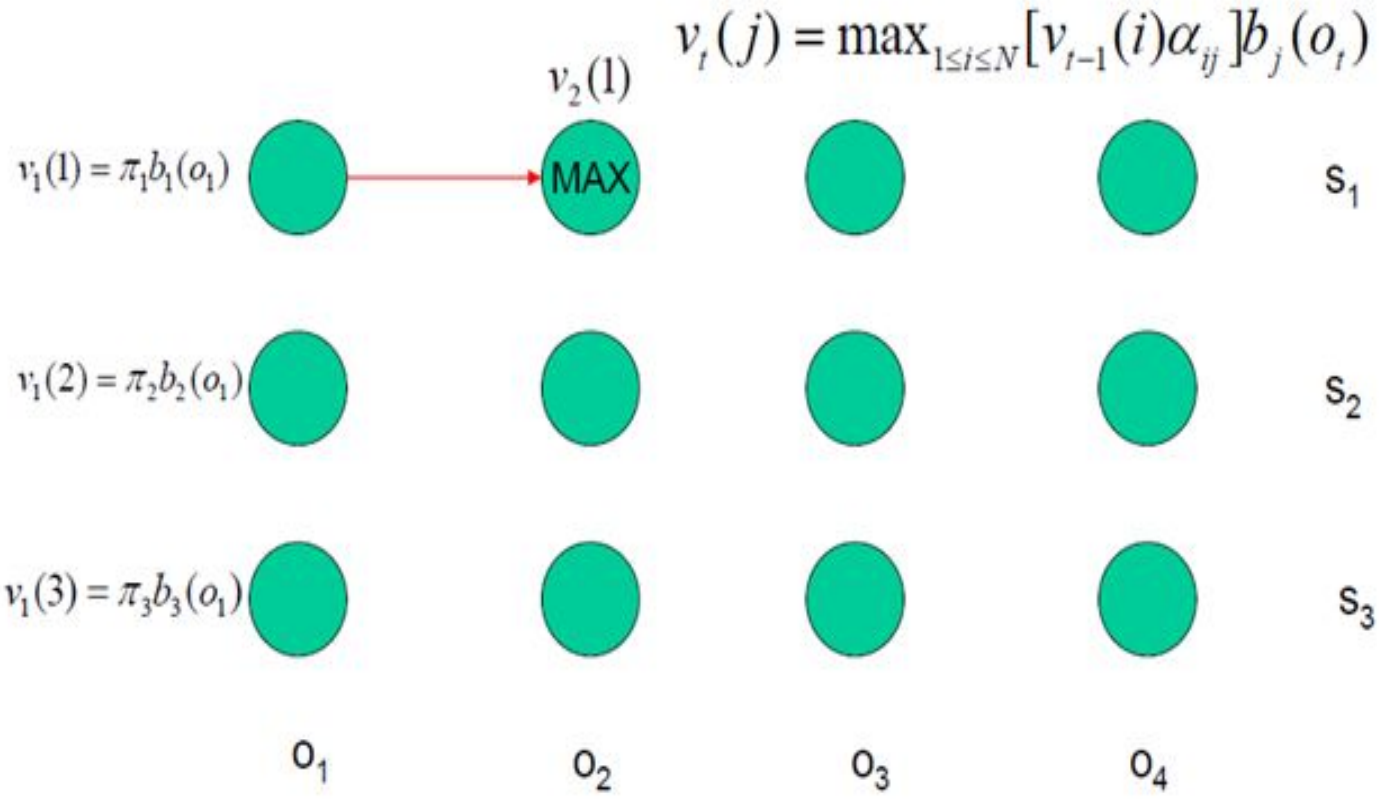
$$P(\hat{X}) = \arg\max_i \delta_i(T)$$

Вычисляем наиболее вероятную последовательность состояний, двигаясь назад

# Viterbi Algorithm

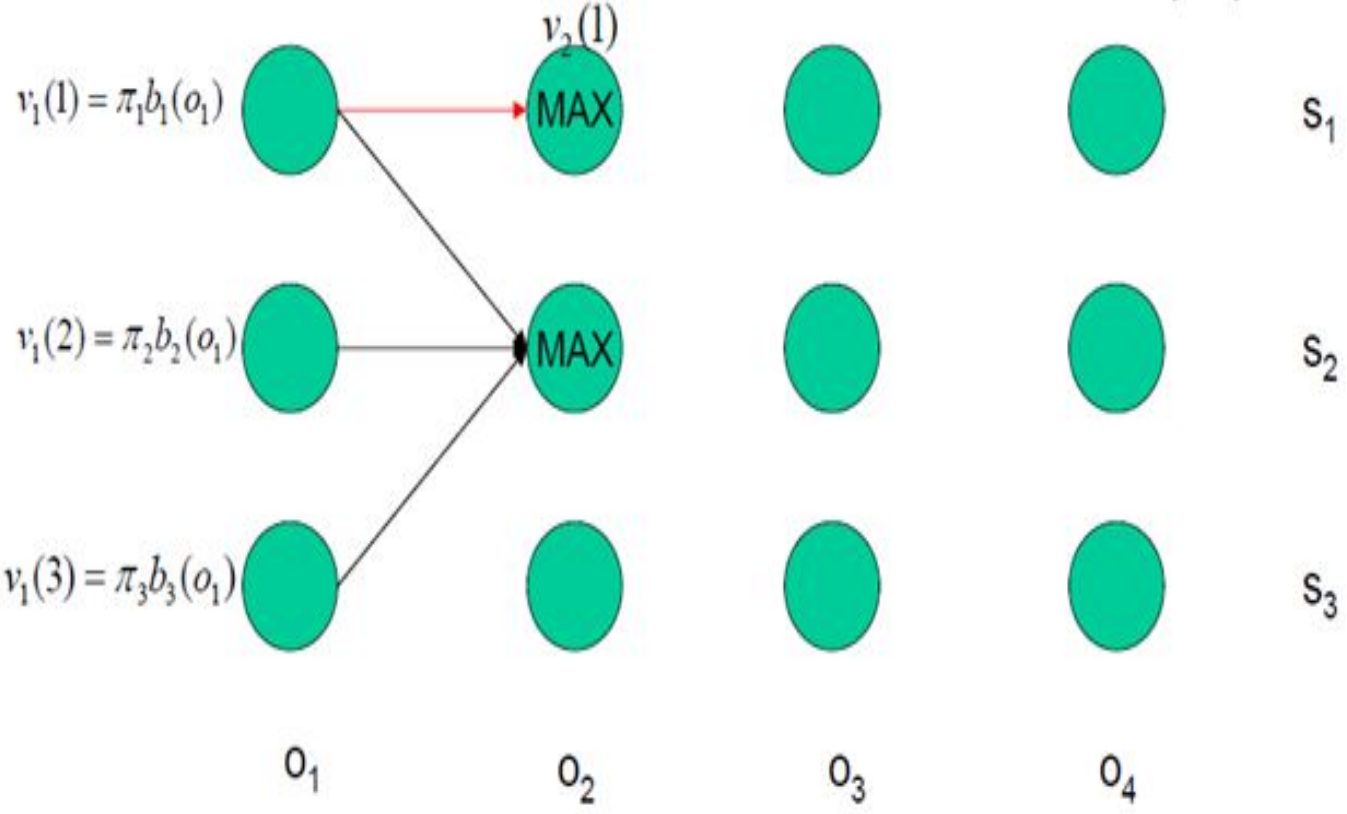


# Viterbi Algorithm

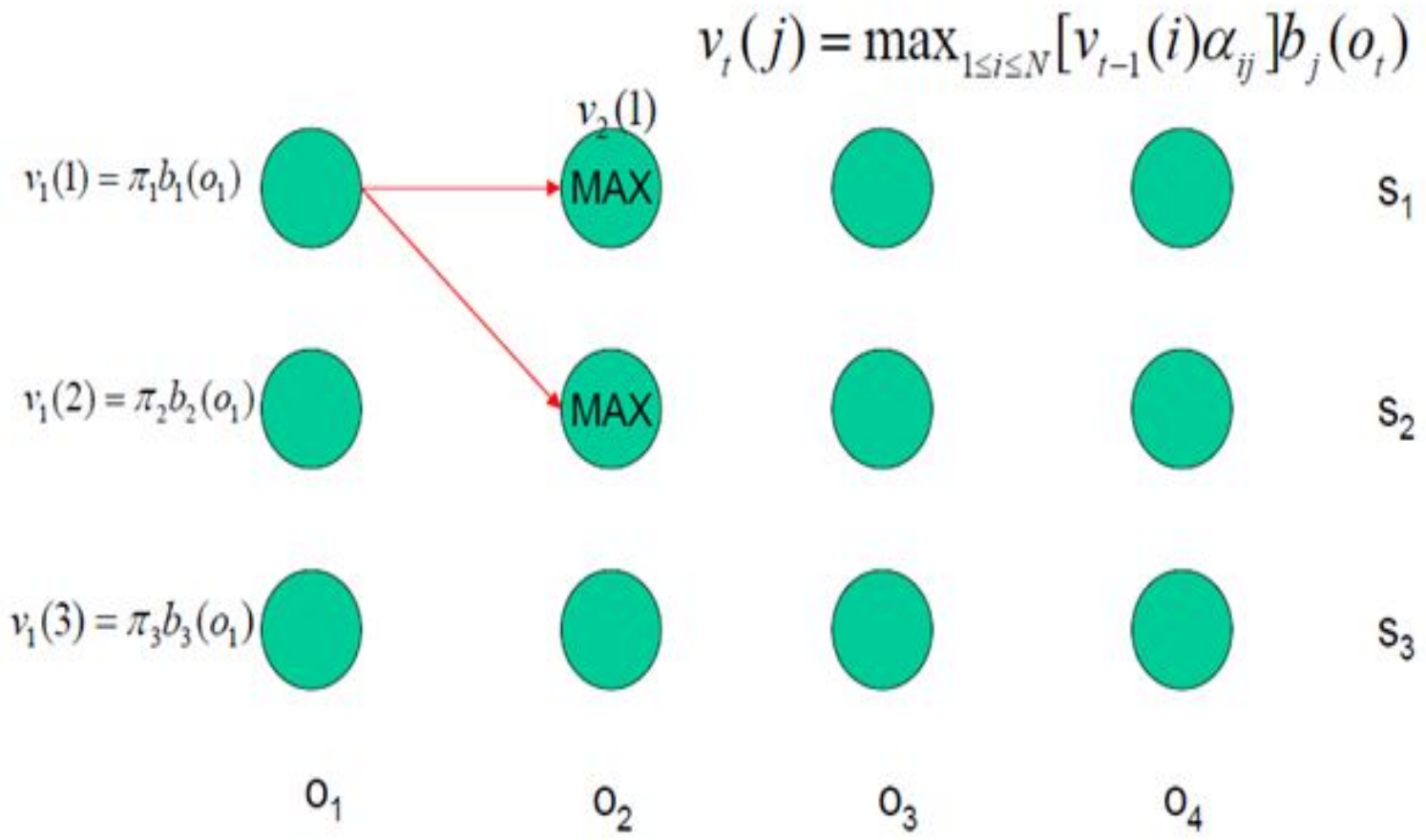


# Viterbi Algorithm

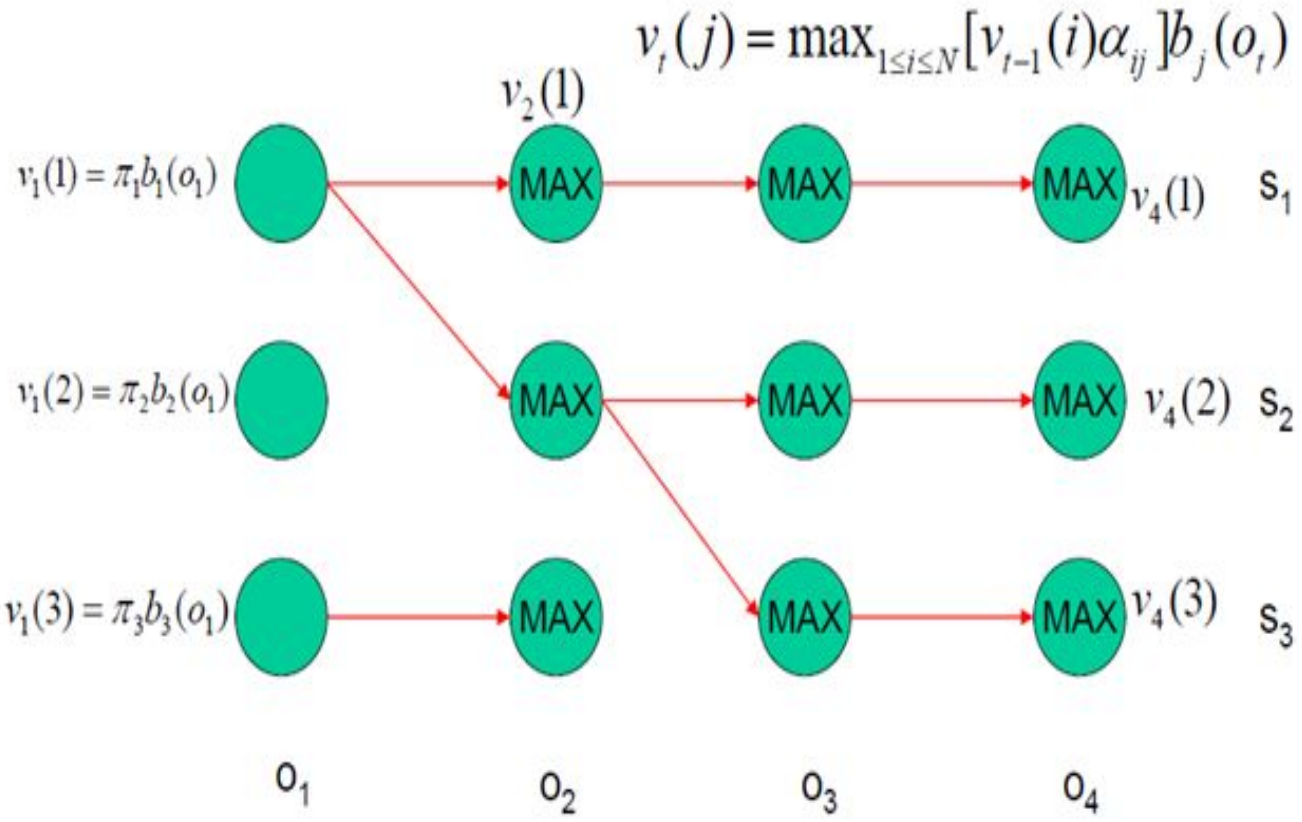
$$v_t(j) = \max_{1 \le i \le N} [v_{t-1}(i)\alpha_{ij}]b_j(o_t)$$



# Viterbi Algorithm

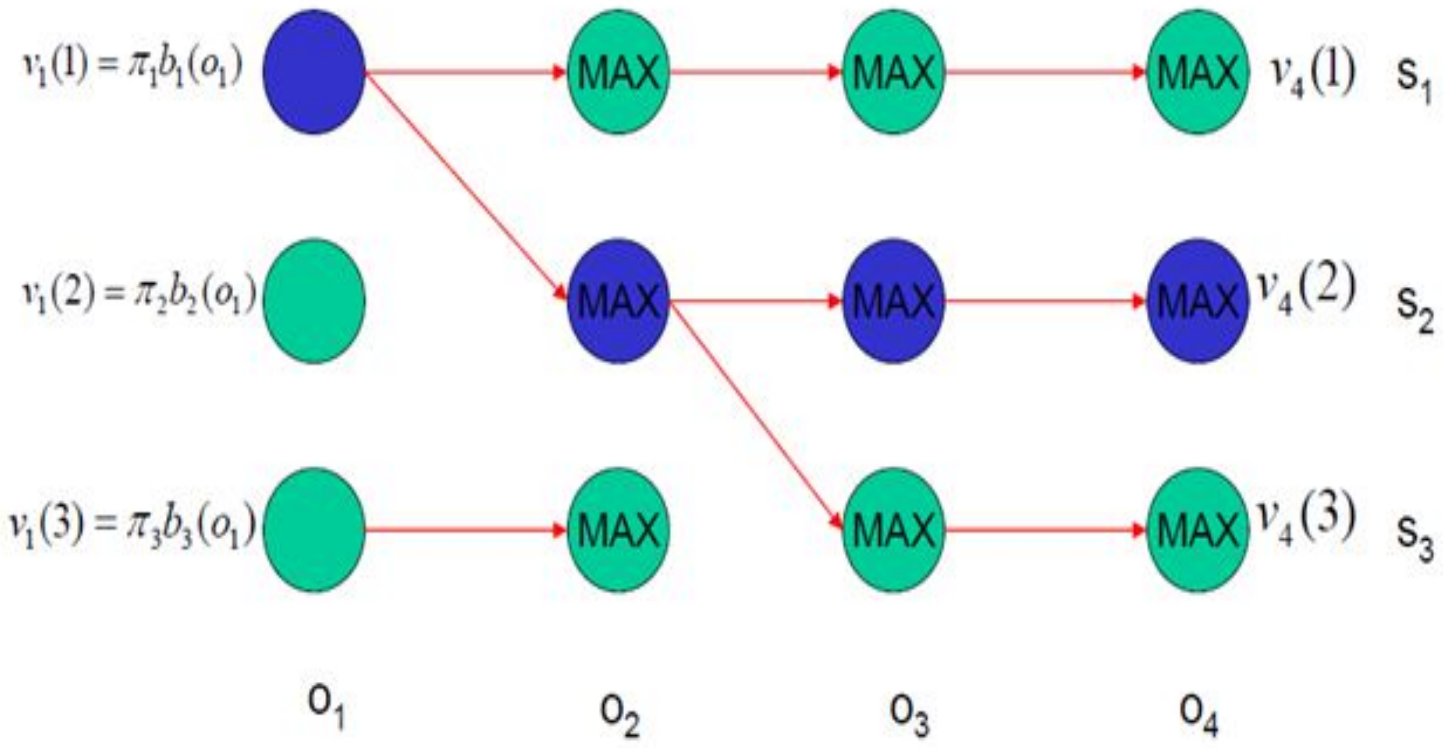


# Viterbi Algorithm



# Viterbi Algorithm

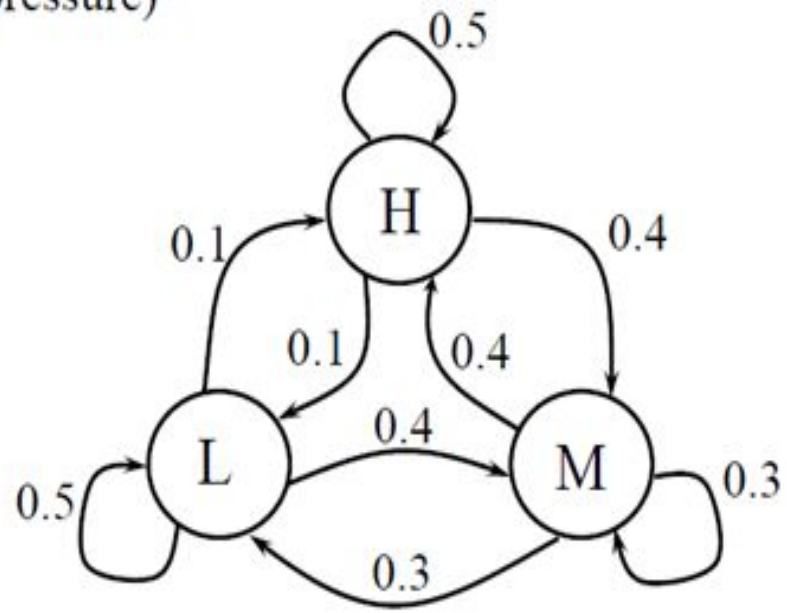
$$\max(v_4(i)) = v_4(2), \quad \arg\max(Q \mid o_1 o_2 o_3 o_4, \Phi) = s_1 s_2 s_2 s_2$$



# Тот же пример для алгоритма Витерби

- Given the following model of the weather:

(states indicate pressure)




| | state M | state H | state L |
|---|---|---|---|
| $P$(sun) | 0.50 | 0.75 | 0.25 |
| $P$(rain) | 0.50 | 0.25 | 0.75 |

$\pi_M = 0.50$
$\pi_H = 0.20$
$\pi_L = 0.30$

# Пример: Алгоритм Витерби

- Same observations: s r r s r

- Compute Viterbi path up to time 5...

$v_1(M)=0.5\cdot0.5$ $\quad$ $v_1(H)=0.2\cdot0.75$ $\quad$ $v_1(L)=0.3\cdot0.25$

$v_1(M)=0.25$ $\quad$ $v_1(H)=0.15$ $\quad$ $v_1(L)=0.075$

$v_2(M) = \max[0.25\cdot0.3, 0.15\cdot0.4, 0.075\cdot0.4]\cdot0.5 = 0.0375$

$v_2(H) = \max[0.25\cdot0.4, 0.15\cdot0.5, 0.075\cdot0.1]\cdot0.25 = 0.0250$

$v_2(L) = \max[0.25\cdot0.3, 0.15\cdot0.1, 0.075\cdot0.5]\cdot0.75 = 0.0563$

$v_3(M) = \max[0.0375\cdot0.3, 0.0250\cdot0.4, 0.0563\cdot0.4]\cdot0.5 = 0.0113$

$v_3(H) = \max[0.0375\cdot0.4, 0.0250\cdot0.5, 0.0563\cdot0.1]\cdot0.25 = 0.0038$

$v_3(L) = \max[0.0375\cdot0.3, 0.0250\cdot0.1, 0.0563\cdot0.5]\cdot0.75 = 0.0211$

# Пример. Алгоритм Витерби

$v_4(M) = \max[0.0113 \cdot 0.3,\ 0.0038 \cdot 0.4,\ 0.0211 \cdot 0.4] \cdot 0.5 = 0.0042$

$v_4(H) = \max[0.0113 \cdot 0.4,\ 0.0038 \cdot 0.5,\ 0.0211 \cdot 0.1] \cdot 0.75 = 0.0034$

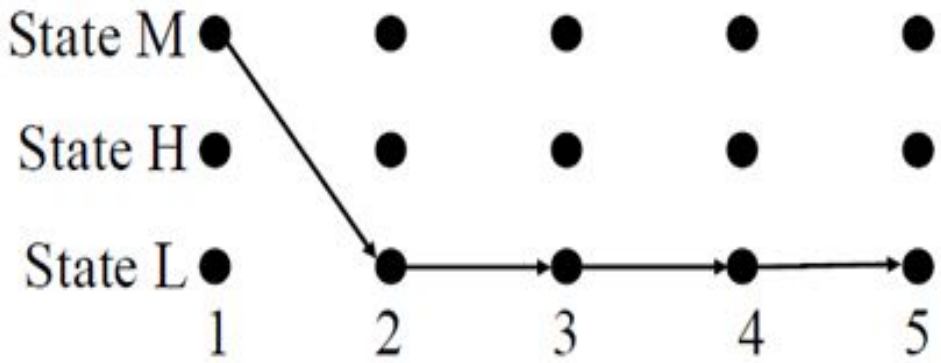
$v_4(L) = \max[0.0113 \cdot 0.3,\ 0.0038 \cdot 0.1,\ 0.0211 \cdot 0.5] \cdot 0.25 = 0.0026$

$v_5(M) = \max[0.0042 \cdot 0.3,\ 0.0034 \cdot 0.4,\ 0.0026 \cdot 0.4] \cdot 0.5 = 0.0007$

$v_5(H) = \max[0.0042 \cdot 0.4,\ 0.0034 \cdot 0.5,\ 0.0026 \cdot 0.1] \cdot 0.25 = 0.0004$

$v_5(L) = \max[0.0042 \cdot 0.3,\ 0.0034 \cdot 0.1,\ 0.0026 \cdot 0.5] \cdot 0.75 = 0.0010$

Maximum state score at time 5 = 0.0010 in state L



# Применение HMM к POS-tagging

- POS-tagging – морфологическая разметка
- HMM tagger: выбирает наиболее вероятную последовательность тегов для каждого предложения
  - Дано предложение $W=w_1, w_2, w_3\ldots, w_n$
  - Вычислить наиболее вероятную последовательность тегов $T=t_1, t_2, \ldots t_n$, которая максимизирует
  - Argmax $P(t_1, t_2, \ldots t_n | w_1, w_2, \ldots w_n)$

# Пример: морфологическая неоднозначность

Example: suppose $w_i$ = race, a verb (VB) or a noun (NN)?
Assume that other mechanism has already done the best tagging to the preceding words.

1) Secretariat/NNP is/VBZ expected/VBN to/TO **race/?** tomorrow
2) People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN the/DT **race/?** for outer space

Bigram $t_i = \arg\max_j P(t_j \mid t_{i-1}) P(w_i \mid t_j)$

Simplify the problem:
to/To race/???
the/DT race/???

→

P(VB|TO) P(race | VB)
P(NN|TO) P(race | NN)

# Откуда взять данные?

- Из корпуса с морфологической разметкой
  - Русский язык:
    - Корпус русского языка
    - Открытый корпус (opencorpora.org)
  - Английский язык
    - Brown corpus
    - Penn tree bank

# Фрагмент морфологической разметки в Национальном корпусе русского языка

- Я сидел на барском сиденье, дышал горячим ветром, бившим в лицо, ощущая в то же время не истребимую никакими сквозняками пыль и легкий запах духов -- катафалк с хорошей скоростью мчался по шоссе на юг. (Ю. Трифонов)

- <s>**Я**{я=S,ед,од=им} **сидел**{сидеть=V,несов=изъяв,прош,ед,муж} **на**{на=PR} **барском**{барский=A=ед,сред,пр} **сиденье**{сиденье=S,сред,неод=ед,пр}, **дышал**{дышать=V,несов=изъяв,прош,ед,муж} **горячим**{горячий=A=ед,муж, твор} **ветром**{ветер=S,муж,неод=ед,твор}, **бившим**{бить=V,несов=прич, прош,ед,муж,твор} **в**{в=PR} **лицо**{лицо=S,сред,неод=ед,вин}, **ощущая** {ощущать=V=несов,деепр,непрош} **в**{в=PR} **то**{тот=A=ед,сред,вин} **же** {же=PART} **время**{время=S,сред,неод=ед,вин} **не**{не=PART} **истребимую** {истребимый=A=ед,жен,вин} **никакими**{никакой=A=мн,твор} **сквозняками** {сквозняк=S,муж,неод=мн,твор} **пыль**{пыль=S,жен,неод,ед=вин} **и**{и=CONJ} **легкий**{легкий=A=ед,муж,вин,неод} **запах**{запах=S,муж,неод=ед,вин}…

# Данные для примера

Look at the Brown and Switchboard corpora

$$P(NN \mid TO) = 0.021$$
$$P(VB \mid TO) = 0.34$$

If we are expecting a verb, how likely it would be “race”

$$P(\text{race} \mid NN) = 0.00041$$
$$P(\text{race} \mid VB) = 0.00003$$

Finally:

$$P(NN \mid TO)\, P(\text{race} \mid NN) = 0.000007$$
$$P(VB \mid TO)\, P(\text{race} \mid VB) = 0.00001$$

# An Example

Example:

*flies like a flower,*

Given

Word sequence $w_1, \ldots, w_N$

POS tags $t_1, \ldots, t_M$, $t_i \in$ [V, N, P, ART]

Find:

Most likely sequence of POS tags $T_1, \ldots, T_N$

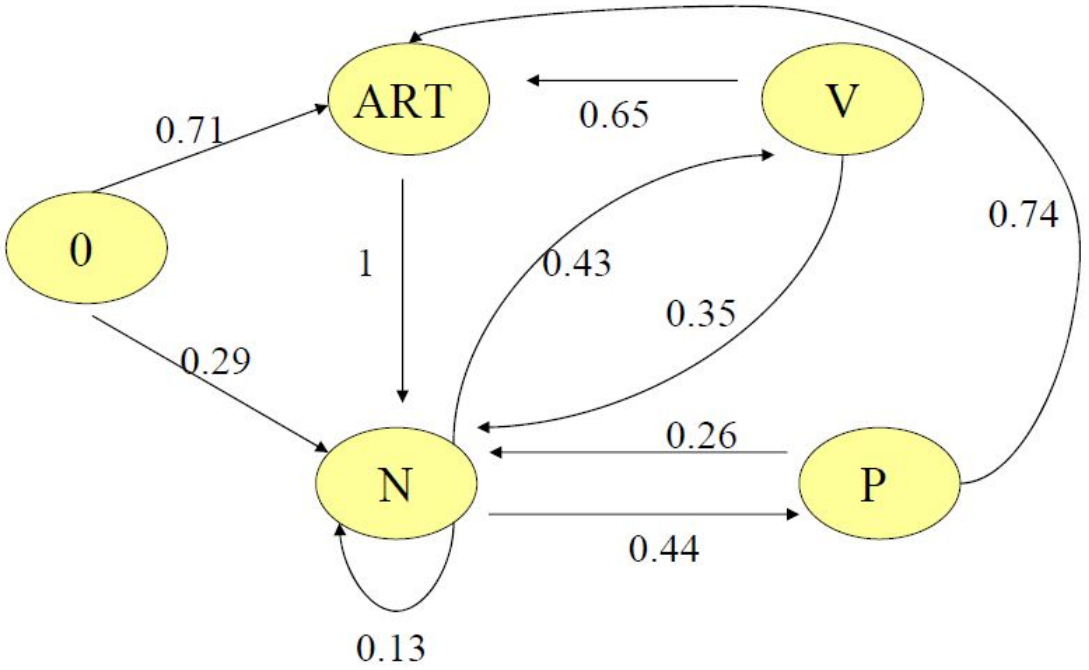
for the word sequence that maximizes:

$P(w_1, \ldots, w_N \mid t_1, \ldots, t_M)$

# Example: Bigram of Tags from a Corpus

| Cat | # at i | Pair | # at i, i+1 | Bigram | Estimate |
|---|---|---|---|---|---|
| 0 | 300 | 0, ART | 213 | Prob(ART\|0) | 0.71 |
| 0 | 300 | 0, N | 87 | Prob(N\|0) | 0.29 |
| ART | 558 | ART, N | 558 | Prob(N \| ART) | 1 |
| N | 833 | N, V | 358 | Prob(V \| N) | 0.43 |
| N | 833 | N, N | 108 | Prob(N \| N) | 0.13 |
| N | 833 | N, P | 366 | Prob(P \| N) | 0.44 |
| V | 300 | V, N | 75 | Porb( N \| V) | 0.35 |
| V | 300 | V, ART | 194 | Prob(ART \| V) | 0.65 |
| P | 307 | P, ART | 226 | Prob (ART \| P) | 0.74 |
| P | 307 | P, N | 81 | Prob (N \| P) | 0.26 |

# A Markov Chain



# Word Counts

| | N | V | ART | P | Total |
|---|---|---|---|---|---|
| flies | 21 | 23 | 0 | 0 | 44 |
| fruit | 49 | 5 | 1 | 0 | 55 |
| like | 10 | 30 | 0 | 21 | 61 |
| a | 1 | 0 | 201 | 0 | 202 |
| the | 1 | 0 | 300 | 2 | 303 |
| flower | 53 | 15 | 0 | 0 | 68 |
| flowers | 42 | 16 | 0 | 0 | 58 |
| bird | 64 | 1 | 0 | 0 | 65 |
| others | 592 | 210 | 56 | 284 | 1142 |
| Total | 833 | 300 | 558 | 307 | 1998 |

$P(the \mid ART)$ ?

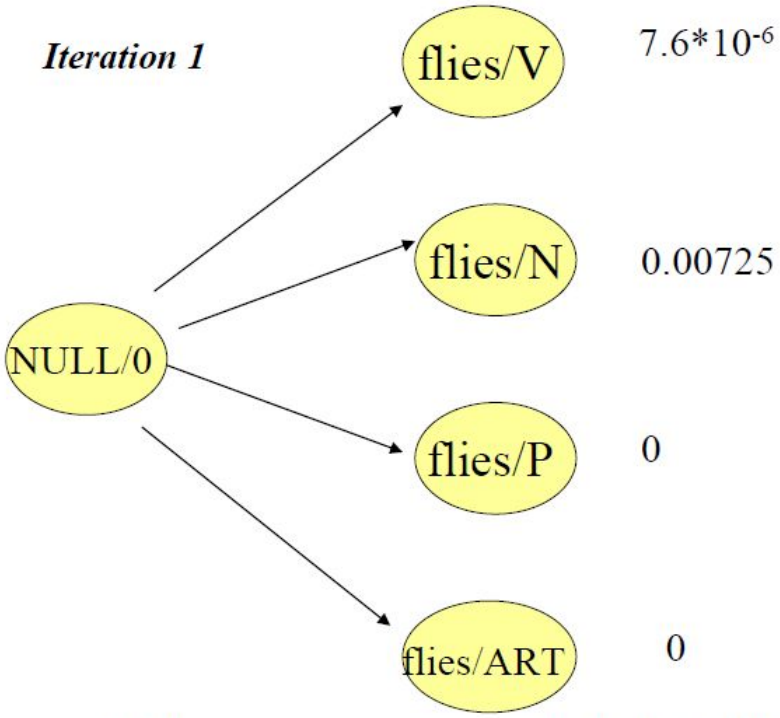
# The Emission Probability

Some examples:

$P(the \mid ART) = 300/558 = 0.54$
$P(flies \mid N) = 0.025$
$P(flies \mid V) = 0.076$
$P(like \mid V) = 0.1$
$P(like \mid P) = 0.068$
$P(like \mid N) = 0.012$

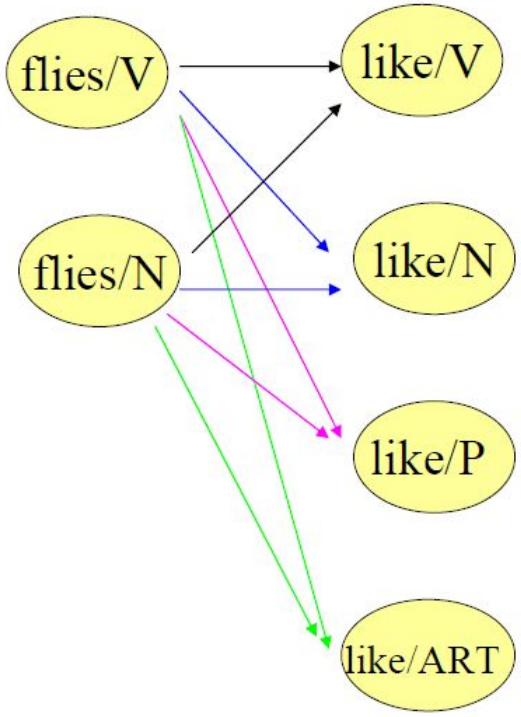
$P(a \mid ART) = 0.360$
$P(a \mid N) = 0.001$
$P(flower \mid N) = 0.063$
$P(flower \mid V) = 0.05$
$P(bird \mid N) = 0.076$

# Viterbi Algorithm - Example



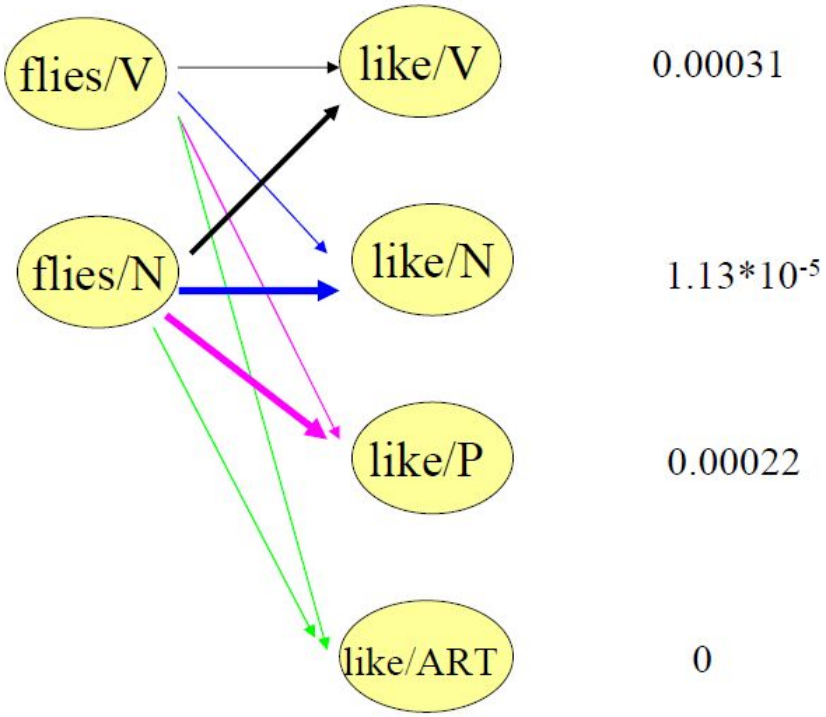
# Viterbi Algorithm - Example

*Iteration 2*



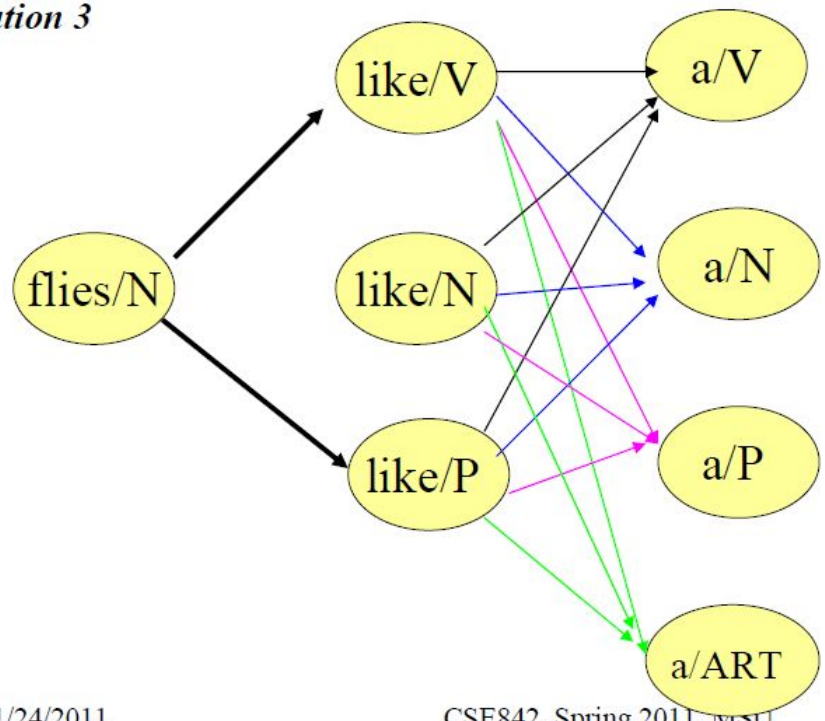
# Viterbi Algorithm - Example

***Iteration 2***



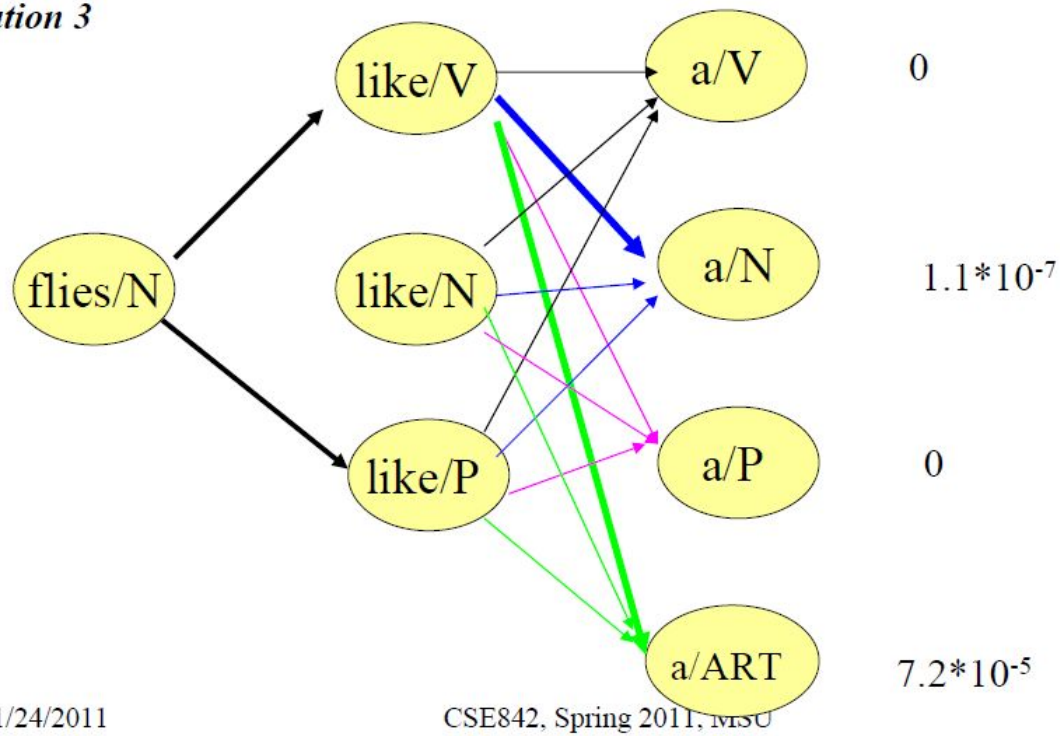
# Viterbi Algorithm - Example

***Iteration 3***

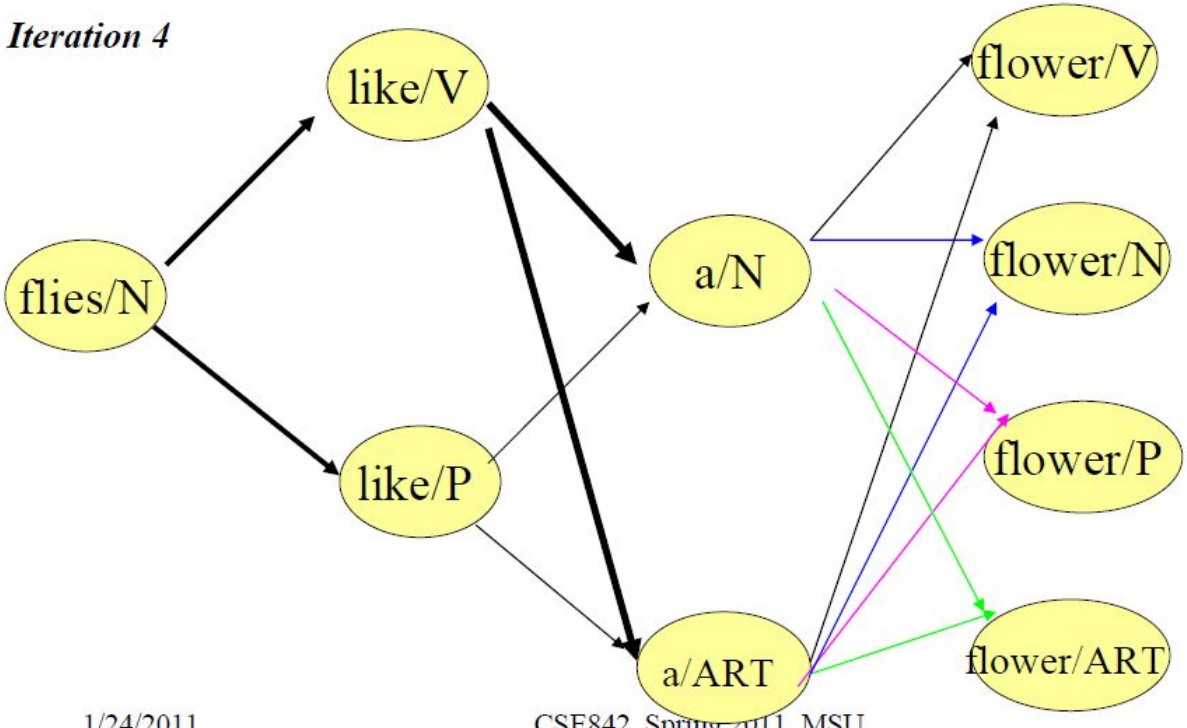


# Viterbi Algorithm - Example

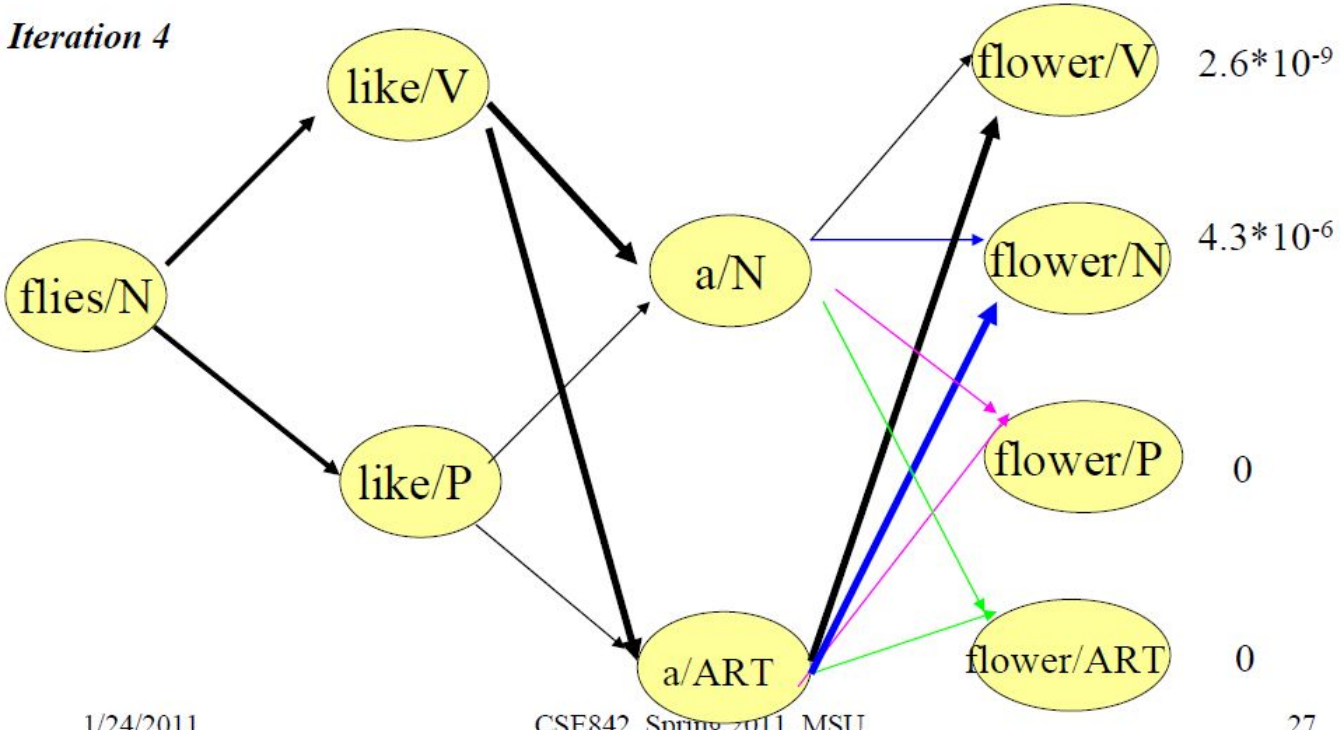
***Iteration 3***



# Viterbi Algorithm - Example



# Viterbi Algorithm - Example



# Лексические вероятности: уточнение

- Мы считали p(w|t)
- Но
  - Слово могло отсутствовать в корпусе или отсутствовать в заданной части речи
  - Не учитывается информация из морфологического словаря
  - Удобнее оценить p (t Iw)

# Лексические вероятности

- 
- $$p(w \mid t) = \frac{p(t \mid w)\,p(w)}{p(t)} \sim \frac{p(t \mid w)}{p(t)}$$

- p(t) – априорная вероятность метки
- p(t|w) – вероятность метки для данного слова
- Можно положить

$$p(t \mid w) = \frac{c(t, w)}{c(w)}$$

- Где c () – количество вхождений
- Как учесть словарь?

# Словарь и лексические вероятности

- Можно считать, что все словарные метки слова w входят в корпус α раз (например, α=0.5)
- Тогда получим:
- $$p(t \mid w) = \frac{c(t,w) + \alpha}{c(w) + \alpha \mid T(w) \mid}$$

- где T(w) – это количество тегов для w
- Для новых несловарных слов p(t|w) считается на основе совокупности признаков (машинное обучение)

# Анализ статистических алгоритмов снятия морфологической омонимии в русском языке

Егор Лакомкин
Иван Пузыревский
Дарья Рыжова
(2013)

# Разрешение морфологической неоднозначности в текстах на английском языке

- Методы:

  Как правило, статистические алгоритмы на основе марковских моделей
- Точность: ~96%

# Особенности английского языка

- Бедная морфология

морфологическая разметка фактически сводится к POS-теггингу

- Фиксированный порядок слов

можно опираться только на локальный контекст слова (ближайших соседей) без учёта дальних зависимостей (т.е. достаточно марковских моделей первого порядка)

# Задача исследования:

Проверить экспериментально, применимы ли статистические алгоритмы, основанные на марковских моделях, к задаче морфологической дизамбигуации текстов на **русском** языке

# Алгоритмы

- Набор скрытых величин Y (*состояний* модели = наборов грамматических тегов); составляют марковскую цепь первого порядка
- Набор наблюдаемых величин X (*наблюдений*) ~ словоформ

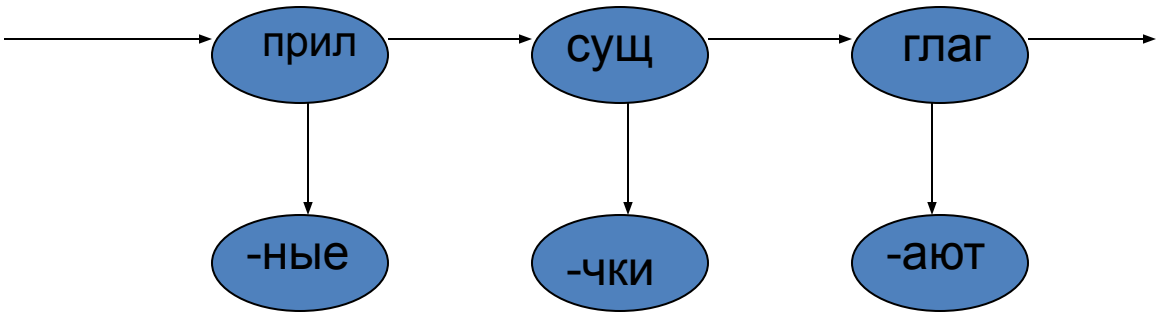
  Словоформы заменяем на 3-буквенные окончания:
  - Сокращаем количество наблюдаемых состояний
  - Практически не теряем полезную информацию (поскольку в РЯ почти вся морфологическая информация сосредоточена в окончании)

# HMM

- Обучение:

Сбор статистик по корпусу:

  - $P(y_i | y_j)$ – матрица переходов
  - $P(x_k | y_i)$ – вероятности наблюдений



# Задача алгоритмов:

Вычисление наиболее вероятной последовательности скрытых величин

# Деление выборки на обучающую и тестирующую:

- **Кросс-валидация** (5 фолдов):
  - Деление выборки на 5 частей:
    4 обучающие + 1 тестирующая
  - 5 серий подсчётов
  - Усреднение результата

# Оценка качества

- Определение верхней и нижней границы:
  - Верхняя граница: процент случаев, когда среди гипотез Mystem'а есть правильная;
  - Нижняя: «частотная снималка» (слову приписывается наиболее частотный вариант разбора, без учёта контекста)

- Качество работы алгоритма (= точность):

  Сравнение с «золотым стандартом» - с эталонным разбором НКРЯ:
  - общая точность
  - точность по знакомым словам
  - точность по незнакомым словам

- Не учитывались:
  - Инициалы, аббревиатуры, цифры;
  - Сложные слова с дефисом (ср. *бело-кремовый*)

# Результаты

| | Части речи | | | Теги (род, число, падеж и др.) | | |
|---|---|---|---|---|---|---|
| | Общ. | Зн. | Незн. | Общ. | Зн. | Незн. |
| Нижн.гр. | .8590 | .8586 | .8885 | .6817 | .6836 | .5525 |
| HMM | .9482 | .9489 | .8996 | .8873 | .8909 | .6550 |
| Верхн.гр. | | .9895 | .9081 | | .9741 | .7017 |

# Выводы работы

- POS-теггинг – на приличном уровне,
- Разрешение неоднозначности по расширенным тегам – довольно низкий уровень точности. Случаи, особенно часто разбираемые ошибочно:
  - Местоимения
  - Имена собственные
  - Субстантивация прилагательных
  - Омонимия падежных форм (номинатив vs. аккузатив)

# Проблемы HMM

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки
  - В русском языке: тег – это сущ. в род. падеже ед. числа
- Ограниченный просмотр состояний – обычно биграммы
- Не учитываются дистантные зависимости
  - Договор о разоружении сторон был подписан
  - Договор – именительный или винительный падеж
- Состояние не зависит от соседних слов
  - Обмануть друга vs. соврать другу

# Как можно изменить процесс расчета переходов между состояниями?

- HMM: учитываются два фактора в простой комбинации
- Для определения вероятности переходов между состояниями нужно: учитывать значительно больше факторов
- Когда нужна комбинация факторов -> машинное обучение

# Задание

Найти оптимальный разбор предложения *Time flies like an arrow*, если матрицы $A$ и $B$ имеют следующий вид:

| | $\langle s \rangle$ | ADJ | N | V | CONJ | DET |
|---|---|---|---|---|---|---|
| $\langle s \rangle$ | 0.0 | 0.2 | 0.2 | 0.1 | 0.01 | 0.4 |
| ADJ | 0.0 | 0.2 | 0.5 | 0.05 | 0.2 | 0.01 |
| N | 0.0 | 0.05 | 0.01 | 0.5 | 0.2 | 0.01 |
| V | 0.0 | 0.1 | 0.2 | 0.01 | 0.2 | 0.3 |
| CONJ | 0.0 | 0.1 | 0.2 | 0.2 | 0.0 | 0.2 |
| DET | 0.0 | 0.3 | 0.7 | 0.0 | 0.0 | 0.0 |

$$p(\textit{time}|\mathrm{N}) = 0.01, \quad p(\textit{time}|\mathrm{V}) = 0.001, \quad p(\textit{time}|\mathrm{ADJ}) = 0.0005,$$
$$p(\textit{flies}|\mathrm{V}) = 0.01, \quad p(\textit{flies}|\mathrm{N}) = 0.0005; \quad p(\textit{like}|\mathrm{CONJ}) = 0.05,$$
$$p(\textit{like}|\mathrm{V}) = 0.02, \quad p(\textit{like}|\mathrm{N}) = 0.001; \quad p(\textit{an}|\mathrm{DET}) = 0.1;$$
$$p(\textit{arrow}|\mathrm{N}) = 0.01, \quad p(\textit{arrow}|\mathrm{ADJ}) = 0.01$$