



# ТЕХНОСФЕРА

## Статистическая обработка текстов

Коллокации, n-граммы, НММ.

Дмитрий Соловьев.

Ведущий разработчик отдела рекомендаций

Москва 2017

# План

- Коллокации. Методы нахождения в текстах
- N – граммы. Общие понятия
- Марковские модели для обработки текстов
- Скрытые Марковские модели. Тегирование

Словосочетание, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого

## **КОЛЛОКАЦИИ**

# Коллокации

- оказать влияние, внести изменения, идет дождь*
  - высокая температура, рост влияния*
- не композиционные выражения.*

*vs. идиомы, которые встречаются редко*

- подложить свинью, биться как рыба об лёд*
  - белая ворона, голодный как волк*
- еще более не композиционные.*

*Коллокации встречаются намного чаще.*

# Признаки коллокации

- Некомпозиционность
  - Смысл коллокации не является композицией смысла частей
- Незаменяемость
  - Нельзя заменять зависимое слово на другое подходящее по смыслу
- Немодифицируемость
  - Компоненты коллокации не получается свободно модифицировать по грамматическим правилам

# Коллокации vs. Термины

Коллокации перекрываются с такими понятиями как:

- Термины
- Технические термины
- Терминологические фразы

# Применение

- Генерация текстов.
- Вычислительная лексикография
- Парсинг
- Корпусные лингвистические исследования

# Коллокации в поиске

- Учет устойчивых словосочетаний в поиске
- Ручной:
  - прирученный; с ручным управлением, неавтоматический; кустарный; лёгкий; послушный, послушливый, покорный, портативный, шелковый, наручный, смирный, мануальный, рукодельный, кроткий, безропотный, покорный
- ‘ручная работа’ = ‘кустарный’ но не ‘лёгкий’
- ‘ручное управление’



# Частотность

- Самый простой способ
- Работает плохо
- Получили одну коллокацию из 20
- Получили 18 803 442 биграмм на 1 миллион документов.

|              | Частота |
|--------------|---------|
| о это        | 41843   |
| один из      | 41694   |
| а также      | 35446   |
| тот что      | 34048   |
| 2015 год     | 33628   |
| в тот        | 32998   |
| что в        | 32007   |
| пресс служба | 30468   |
| и в          | 27138   |
| в это        | 26101   |
| не быть      | 24956   |
| отметить что | 24088   |
| при это      | 22607   |
| из за        | 22398   |
| о тот        | 21240   |

# Частотность + Эвристика

- Учитываем части речи
  - AN – учебный год
  - NN – пресс служба
  - AAN - дискретная случайная величина
  - ANN - эмпирическая функция распределения
  - NAN -
  - NNN -
  - NPN -

|                      | Частота |
|----------------------|---------|
| пресс служба         | 129071  |
| тот число            | 69728   |
| уголовный дело       | 57108   |
| риа новость          | 54353   |
| миллион рубль        | 52868   |
| настоящее время      | 44442   |
| миллиард рубль       | 44368   |
| прошное год          | 43679   |
| vladimir putin       | 41806   |
| российский федерация | 37965   |
| такой образ          | 36246   |
| премьер министр      | 33102   |
| тысяча рубль         | 29209   |
| санкт петербург      | 28120   |
| данный момент        | 26737   |
| главный тренер       | 26123   |

# Среднее и дисперсия

- Поиск по частотным вхождениям хорош для фиксированных фраз.
- ‘постучаться в дверь’, но
  - ‘Она постучалась в дверь’
  - ‘Они постучались в эту деревянную дверь’
  - ‘Человек постучался в металлическую дверь’

# Среднее и дисперсия

- Используем окно 3-4 слова
  - Экономика балансирует на грани коллапса

*Экономика балансирует / экономика на / экономика грани  
балансирует на / балансирует грани / балансирует коллапса  
на грани / на коллапса  
грани коллапса*

Используем окно в 3 слова

# Среднее и дисперсия

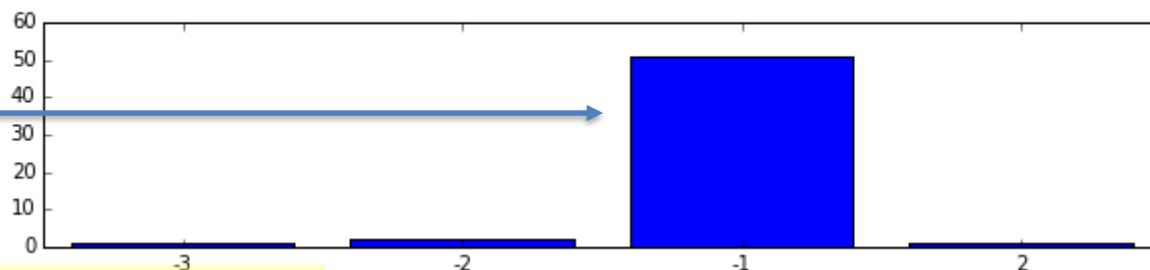
- Пример: 'постучать ... дверь'
- $\mu = (2+4+3) / 3 = 3$
- $\sigma^2 = \frac{\sum_{i=1}^n (d_i - \mu)^2}{n-1} = \frac{1}{2} (2 - 3)^2 + (4 - 3)^2 + (3 - 3)^2 = 1$
- $\sigma = 1$
- Нулевая дисперсия говорит, что эти слова всегда встречаются на одном расстоянии.

# Среднее и дисперсия

Маленькие дисперсия и среднее

сильная волатильность

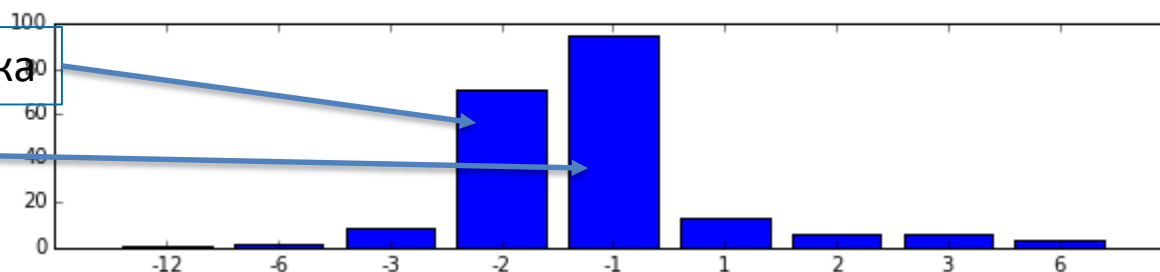
mean: -1.02 std: 0.52/ (волатильность сильный)



mean: -1.10 std: 1.74/ (поддержка сильный)

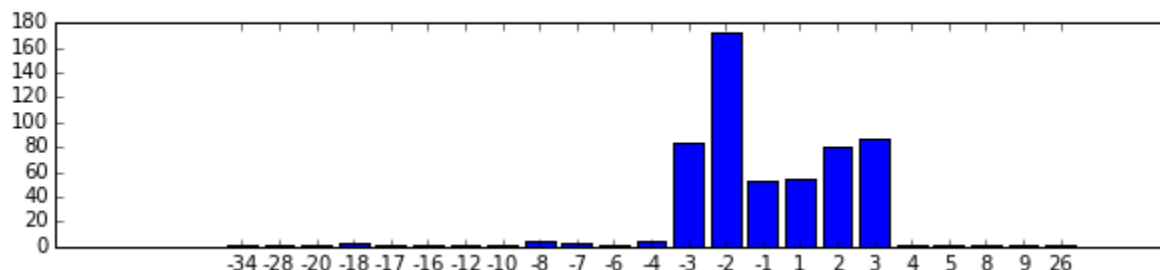
сильная медийная поддержка

сильная поддержка



mean: -0.75 std: 3.84/ (для сильный)

Нет интересных коллокаций



# Среднее и дисперсия

Определяем  
правдоподобие  
коллокаций через  
среднее и  
дисперсию

|                      | Частота | Mean      | Std      |
|----------------------|---------|-----------|----------|
| пресс служба         | 129071  | -1.064374 | 1.230774 |
| тот число            | 69728   | -0.813762 | 2.234214 |
| уголовный дело       | 57108   | -0.969454 | 2.317335 |
| риа новость          | 54353   | -1.001122 | 0.183002 |
| миллион рубль        | 52868   | -0.973434 | 2.168581 |
| настоящее время      | 44442   | -1.121742 | 1.469904 |
| миллиард рубль       | 44368   | -0.692835 | 3.690560 |
| прошрое год          | 43679   | -1.791702 | 4.549470 |
| vladimir putin       | 41806   | -1.007878 | 2.046794 |
| российский федерация | 37965   | -0.702253 | 2.904589 |
| такой образ          | 36246   | -0.914430 | 1.762869 |
| премьер министр      | 33102   | -1.344328 | 2.6369   |

# Проверка гипотез

- Можем ли мы сказать, что биграмма “новая компания” является коллокацией?
- Встречаются ли два слова со смыслом или случайно?
- Есть ли способ отделить случайное событие от неслучайного?



# Проверка гипотез

- Классическая задача из статистики:
  - Формулируем нулевую гипотезу  $H_0$ :
    - Между словами нет никакой связи, только случайная встречаемость
    - Вычисляем вероятность  $p$  с учетом истинности  $H_0$
    - Отвергаем  $H_0$  если  $p$  ниже некоторого порога (0.01, 0.001... )
    - В противном случае принимаем гипотезу  $H_0$

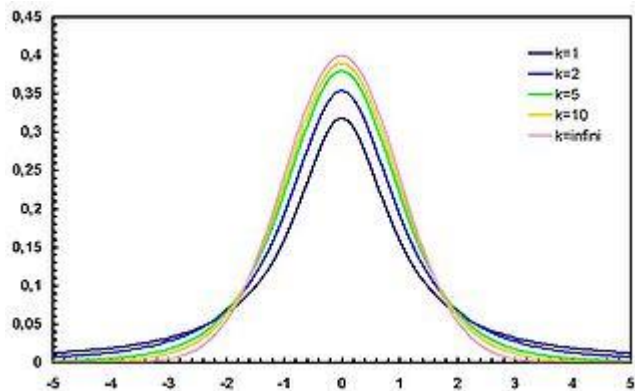
# Проверка гипотез

- Как использовать метод для проверки коллокаций?
  - $H_0 = true$  если два слова ( $w_1$  ;  $w_2$ ) не формируют коллокацию =>
  - Предполагаем что  $w_1$  и  $w_2$  генерируются независимо друг от друга
  - $P(w_1, w_2) = P(w_1)P(w_2)$
  - Простая, не совсем точная модель подходит для решения задачи

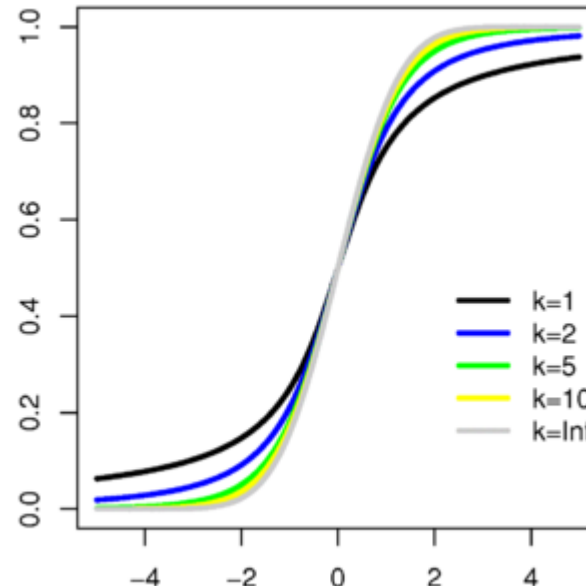
# Распределение Стюдента

- Пусть  $Y_0, Y_1, \dots, Y_n$  - независимые стандартные нормальные случайные величины
- $Y_i \sim N(0, 1) \quad i = 0, \dots, n$
- Распределение случайной величины  $t$ :
- $t = \frac{Y_0}{\sqrt{\frac{\sum_{i=1}^n Y_i^2}{n}}}$  - распределение Стюдента с  $n$  – степенями свободы

# Распределение Стьюдента



Плотность вероятности



Функция распределения

Для этого распределения  
рассчитаны таблицы  
значений квантилей для  
разных степеней свобод.

# Проверка гипотез

- Критерий Стьюдента (t - критерий)
- Допущение – примеры отбираются из нормального распределения
- Нулевая гипотеза – пример взят из выборки с мат ожиданием  $\mu$

- $$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

|           |                               |
|-----------|-------------------------------|
| $\bar{x}$ | среднее выборки               |
| $s^2$     | дисперсия выборки             |
| $N$       | количество примеров в выборке |

# t - критерий

- Если  $t$  достаточно велико, то  $H_0$  отвергается
- Насколько большим должно быть  $t$  проверяется по таблицам  $t$  распределения

|                             |      | Достоверность |       |       |       |       |        |
|-----------------------------|------|---------------|-------|-------|-------|-------|--------|
|                             |      | $P$           | 0.05  | 0.025 | 0.01  | 0.005 | 0.0005 |
|                             |      | C             | 90%   | 95%   | 98%   | 99%   | 99.8%  |
| Количество степеней свободы | d.f. | 1             | 6.314 | 12.71 | 31.82 | 63.66 | 318.3  |
|                             |      | 10            | 1.812 | 2.228 | 2.764 | 3.169 | 4.587  |
|                             |      | 20            | 1.725 | 2.086 | 2.528 | 2.845 | 3.552  |
|                             | (z)  | $\infty$      | 1.645 | 1.960 | 2.326 | 2.576 | 3.091  |

# t – критерий. Пример

- $H_0$  гипотеза: средняя высота популяции мужчин 158 см
- Мы взяли выборку 200 человек :

$$- \bar{x} = 169, s^2 = 2600$$

$$- t = \frac{169 - 158}{\sqrt{2600/200}} \approx 3,05$$

| P    |          | 0.05  | 0.025 | 0.01  | 0.005 | 0.001 | 0.0005 |
|------|----------|-------|-------|-------|-------|-------|--------|
| C    |          | 90%   | 95%   | 98%   | 99%   | 99.8% | 99.9%  |
| d.f. | 1        | 6.314 | 12.71 | 31.82 | 63.66 | 318.3 | 636.6  |
|      | 10       | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587  |
|      | 20       | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850  |
| (z)  | $\infty$ | 1.645 | 1.960 | 2.326 | 2.576 | 3.091 | 3.291  |

– при уровне достоверности  $\alpha = 0.005$  находим значение: 2,576.

–  $3,05 > 2,576$  с 99% вероятностью отвергаем гипотезу

# t – критерий в коллокациях

- Проверим гипотезу:
  - “*миллион рублей*” – не коллокация
    - частота слов: 'миллион' - 215131 , 'рубль': 268089
    - всего слов: 186419524
    - $P(\text{миллион}) = 215131 / 186419524 = 0.0012$
    - $P(\text{рубль}) = 268089 / 186419524 = 0.0014$
    - $H_0 : P(\text{миллион рублей}) = P(\text{миллион})P(\text{рубль}) = 0.0012 * 0.0014 = 0.0000017$

(пример из тетрадки)



# t – критерий. Пример №2

- Если  $H_0 = true \Rightarrow$ 
  - процесс генерации биграммы имеет распределение бернулли
    - 1 – ‘миллион рублей’ присутствует
    - 0 - ‘миллион рублей’ отсутствует
    - С вероятностью  $P = 0.0000017$
    - $\mu = p = 0.0000017$ ;  $\sigma^2 = p(1 - p) = 0.00000166 \approx p$
    - это наше ожидаемое значение

(пример из тетрадки)

## t – критерий. Пример №2

- Для биграммы получим реальное среднее

$$- \bar{x} = \frac{52868}{186419524} = 0.0002836$$

– Проводим t – тест:

$$- t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} = 228.58 \gg 2.576 \text{ при } \alpha = 0.005$$

– мы можем отвергнуть гипотезу о независимости слов.

*(пример из тетрадки)*

# $t$ – критерий. Пример №3

- подсчитаем достоверность биграммы 'новая компания'
- см. тетрадку:  $t = 3.797 > 2.576$  при достоверности 99%
- гипотеза может быть отвергнута

*(пример из тетрадки)*

# Проверка гипотез

- Критерий Пирсона.  $\chi^2$  - критерий
  - Критерий Стьюдента подразумевает нормальное распределение
  - Критерий Пирсона не делает таких предположений.
  - Критерий сравнивает частоты наблюдаемые с ожидаемыми в случае независимости событий.

# Критерий Пирсона

|                            | $w_1 = \text{НОВЫЙ}$     | $w_1 \neg \text{НОВЫЙ}$      |
|----------------------------|--------------------------|------------------------------|
| $w_2 = \text{компания}$    | 335<br>(новая компания)  | 233264<br>(старая компания)  |
| $w_2 \neg \text{компания}$ | 211541<br>(новая машина) | 185974049<br>(старая машина) |

- $C(\text{НОВЫЙ}) - 211876$
- $C(\text{компания}) - 233599$
- $C(\text{новая компания}) - 335$
- Всего токенов - 186419524

# Критерий Пирсона

- Суммирует разницу между наблюдаемым и ожидаемым значением

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2$$

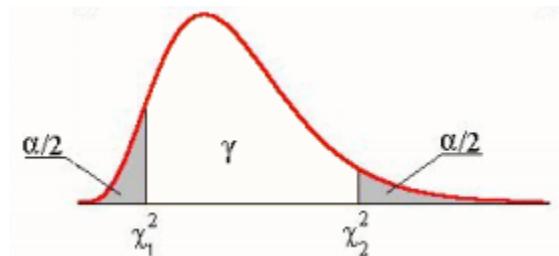
i    итерация по строкам  
 j    итерация по столбцам  
 E    ожидаемое значение  
 O    наблюдаемое значение

- $\chi^2$  асимптотически сходится к  $\chi^2$

$\chi_1^2 \leq X^2 \leq \chi_2^2$  - гипотеза  $H_0$  выполняется

$\chi_1^2 \geq X^2$  -

$\chi_2^2 \geq X^2$  - гипотеза не выполняется



# Критерий Пирсона. Пример

- Ожидаемая частота биграммы 'новая компания':

|                            | $w_1 = \text{новый}$     | $w_1 \neg \text{новый}$      |
|----------------------------|--------------------------|------------------------------|
| $w_2 = \text{компания}$    | 335<br>(новая компания)  | 233264<br>(старая компания)  |
| $w_2 \neg \text{компания}$ | 211541<br>(новая машина) | 185974049<br>(старая машина) |

- $$\frac{P_{\text{компания}}}{N} \times \frac{P_{\text{новая}}}{N} \times N = 265.50$$
- Для таблицы размеров 2x2:

$$-\frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} = 18.2375$$

# Критерий Пирсона. Пример

|                 |          | Вероятность отсечения |        |       |       |       |       |       |
|-----------------|----------|-----------------------|--------|-------|-------|-------|-------|-------|
| Степени свободы | <i>p</i> | 0.99                  | 0.95   | 0.10  | 0.05  | 0.01  | 0.005 | 0.001 |
|                 | d.f. 1   | 0.00016               | 0.0039 | 2.71  | 3.84  | 6.63  | 7.88  | 10.83 |
|                 | 2        | 0.020                 | 0.10   | 4.60  | 5.99  | 9.21  | 10.60 | 13.82 |
|                 | 3        | 0.115                 | 0.35   | 6.25  | 7.81  | 11.34 | 12.84 | 16.27 |
|                 | 4        | 0.297                 | 0.71   | 7.78  | 9.49  | 13.28 | 14.86 | 18.47 |
|                 | 100      | 70.06                 | 77.93  | 118.5 | 124.3 | 135.8 | 140.2 | 149.4 |

- Имея одну степень свободы для таблицы 2x2, по всем уровням вероятности гипотеза  $H_0$  не может быть отброшена



# $\chi^2$ для перевода

- Считаем таблички совместной встречаемости для двуязычных текстов

|                     | <i>cow</i> | $\neg$ <i>cow</i> |
|---------------------|------------|-------------------|
| <i>vache</i>        | 59         | 6                 |
| $\neg$ <i>vache</i> | 8          | 570934            |

$$\chi^2 = 456400.$$

- Предполагаем, что *vache* хороший перевод слова *cow*

# Проверка гипотез

- Критерий отношения правдоподобия
- Лучше подходит для разряженных данных чем  $\chi^2$
- Дает более понятную интерпретацию результата, т.е. во сколько раз одна гипотеза лучше чем другая.

# Отношение правдоподобия

- Проверяем:
- H1:  $P(w_2 | w_1) = p = P(w_2 | \neg w_1)$  – гипотезу о независимости
- H2:  $P(w_2 | w_1) = p_1 \neq p_2 = P(w_2 | \neg w_1)$  гипотезу о зависимости

# Отношение правдоподобия

- $p = \frac{c_2}{N}$  ;  $p_1 = \frac{c_{12}}{c_1}$  ;  $p_2 = \frac{c_2 - c_{12}}{N - c_1}$
- $p$  - вероятность;  $c$  – частота встречаемости слов
- Предполагаем биномиальное распределение:  
–  $b(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$

# Отношение правдоподобия

- Тогда для гипотез
  - $L(H_1) = b(c_{12}; c_1, p)b(c_2 - c_{12}; N - c_1, p)$
  - $L(H_2) = b(c_{12}; c_1, p_1)b(c_2 - c_{12}; N - c_1, p_2)$

$$P(w_2|w_1)$$

$$P(w_2|\neg w_1)$$

$c_{12}$  вне  $c_1$  и биграмма  $w_1w_2$

$c_2 - c_{12}$  вне  $N - c_1$  и  
биграмма  $\neg w_1w_2$

$H_1$

$$p = \frac{c_2}{N}$$

$$p = \frac{c_2}{N}$$

$$b(c_{12}; c_1, p)$$

$$b(c_2 - c_{12}; N - c_1, p)$$

$H_2$

$$p_1 = \frac{c_{12}}{c_1}$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

$$b(c_{12}; c_1, p_1)$$

$$b(c_2 - c_{12}; N - c_1, p_2)$$

# Отношение правдоподобия

- Выражаем отношение правдоподобия:

$$\begin{aligned}\log \lambda &= \log \frac{L(H_1)}{L(H_2)} \\ &= \log \frac{b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c_1, p_2)} \\ &= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\ &\quad - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)\end{aligned}$$

- $-2 \log \lambda \sim \chi^2 \Rightarrow$  можно пользоваться таблицами для проверки гипотезы  $H_0$ .

# Отношение правдоподобия

пример из тетрадки

|                      | $-2\log \lambda$ | C1    | C2     | C12 |
|----------------------|------------------|-------|--------|-----|
| мощный россия        | 66.800607        | 14709 | 513837 | 2   |
| мощный землетрясение | 58.521963        | 14709 | 5490   | 543 |
| мощный взрыв         | 44.635622        | 14709 | 20928  | 930 |
| мощный работа        | 19.562579        | 14709 | 242103 | 4   |
| мощный город         | 15.886387        | 14709 | 212631 | 4   |
| мощный решение       | 15.403840        | 14709 | 185295 | 3   |
| мощный матч          | 14.409602        | 14709 | 222457 | 5   |
| мощный образ         | 4.061203         | 14709 | 69798  | 2   |
| мощный граница       | 3.767401         | 14709 | 67112  | 2   |
| мощный орган         | 3.509477         | 14709 | 82553  | 3   |
| мощный позиция       | 3.261025         | 14709 | 62387  | 2   |
| мощный тайфун        | 3.239385         | 14709 | 3196   | 92  |
| мощный ливень        | 3.051796         | 14709 | 1551   | 62  |

С достоверностью 0.001, можно отвергнуть  $H_0$


С достоверностью 0.005, можно принять  $H_0$

# N-ГРАММЫ



# N-грамм модель

- Вероятность появления следующего слова зависит от последовательности предшествующих:  $p(w_n | w_1, w_2, \dots, w_{n-1})$



Не делаем предположения о порядке следования

N – грамм


- 1 - униграмма (unigramm)
- 2 - биграмма (bigramm)
- 3 - триграмма ( - )

# Как это работает:

... большую зеленую {  
таблетку  
лягушку

# Как это работает:

... съел большую зеленую



таблетку

лягушку

# Словарь из 20К слов

| Модель      | Количество параметров                           |
|-------------|---|
| Биграммная  | $20\,000 \times 19\,999 = 400$ милл.            |
| Триграммная | $20\,000^2 \times 19\,999 = 8$ трилл.           |
| 4х граммная | $20\,000^3 \times 19\,999 = 1,6 \times 10^{17}$ |

Как быть:

- Уменьшать  $n$
- Делать классы эквивалентности (стемминг, синонимы ... )

# Статистическая оценка модели

- Оцениваем условную вероятность:
- $$p(w_n | w_1, w_2 \dots w_{n-1}) = \frac{p(w_1, w_2, \dots, w_n)}{p(w_1, w_2, \dots, w_{n-1})}$$
- Нужно оценить:  $p(w_1, w_2, \dots, w_n)$

# Пример: 3-gramm

| $(w_1, w_2)$    | $(w_3)$  | $C(w_1, w_2, w_3)$ | $P(w_1, w_2, w_3)$ |
|-----------------|----------|--------------------|--------------------|
| большую зеленую |          | N= 10              | -                  |
| большую зеленую | лягушку  | 8                  | 0.8                |
| большую зеленую | таблетку | 1                  | 0.1                |
| большую зеленую | сумку    | 1                  | 0.1                |

# Maximum Likelihood Estimator

- Оценка через относительную частоту – MLE

$$p_{MLE}(w_1, w_2, \dots, w_n) = \frac{C(w_1, w_2, \dots, w_n)}{N}$$

- $N$  – количество встречаемости  $n-1$  - грамм
- тогда

$$p_{MLE}(w_n \mid w_1, w_2, \dots, w_{n-1}) = \frac{C(w_1, w_2, \dots, w_n)}{C(w_1, w_2, \dots, w_{n-1})}$$

# Оценка фразы

- Оценим фразу

*“Президент США Барак Обама решил сняться в телепередаче с Беар Гриллсом.”*

моделями:

- Unigram
- Bigram
- Trigram

*пример из тетрадки*



# Пример: Оценка модели unigram

|      | президент | сша     | барак    | обама    | решить   | сняться  | в        | телепередача | с        | бears    | грилсом  |
|------|-----------|---------|----------|----------|----------|----------|----------|--------------|----------|----------|----------|
| idx  | 76        | 116     | 3395     | 1799     | 328      | 4657     | 0        | 15613        | 3        | 14581    | 21467    |
| prob | 0.001167  | 0.00084 | 0.000037 | 0.000081 | 0.000395 | 0.000024 | 0.043321 | 0.000004     | 0.011905 | 0.000004 | 0.000002 |

- Хорошие вероятности
- Плохая позиция

*пример из тетрадки*

Топовые слова по вероятностям

- в - 0.0433
- и - 0.0245
- на - 0.0194

# Пример: Оценка модели bigram

|      | президент<br>сша | сша барак | барак обама | обама решить | решить<br>сняться | сняться в | в<br>телепередача | телепередача<br>с | с беар   | беар<br>грилсом |
|------|------------------|-----------|-------------|--------------|-------------------|-----------|-------------------|-------------------|----------|-----------------|
| num  | 4                | 2         | 0           | 66           | 264               | 0         | 3735              | 16                | 2105     | 1               |
| prob | 0.034506         | 0.022638  | 0.931526    | 0.002606     | 0.000671          | 0.451685  | 0.000018          | 0.014085          | 0.000062 | 0.316456        |

## Топ вероятностей биграмм для слова президент

- президент россия 0.076685584563
- президент украина 0.0706923950057
- президент рф 0.0625652667423
- президент рфс 0.0429511918275

*пример из тетрадки*

# Пример: Оценка модели 3-gram

- Выполните самостоятельно.

## В итоге

- Unigram – полностью игнорирует контекст, но это не бесполезно для общих слов
- Bigram - использует предыдущее слово для оценки вероятности, получаем лучшую модель.
- 3-gram модель должна работать хорошо, но ... Появляется очень много дырок в оценке вероятности.
- ...

# Как бороться с дырками

- В сложных моделях 3-gram и выше – можно комбинировать эстиматоры:
  - Мало данных – снижаем  $n$
  - Много данных – повышаем  $n$
- Для случая с биграммами
  - Сглаживание
- Сглаживание можно использовать для  $n$ -gram любых порядков

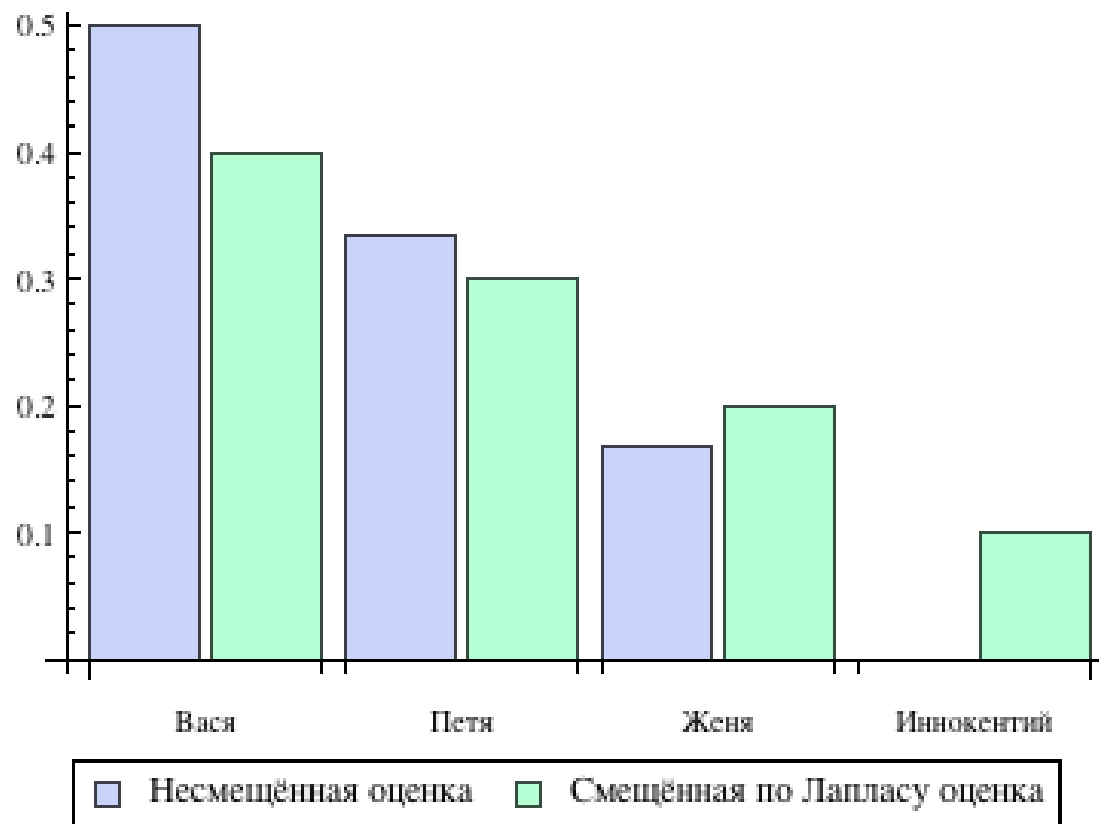
# Сглаживание Лапласа (adding one)

- Самый простой вид сглаживания
- $$p_{lap}(w_1 \dots w_n) = \frac{C(w_1 \dots w_n) + 1}{C(w_1 \dots w_{n-1}) + B}$$
- B – размер словаря
- Провоцирует сильную погрешность
- Иногда несглаженная модель показывает лучшие результаты

# Пример

| Имя  | Частота |
|------|---------|
| Вася | 3       |
| Петя | 2       |
| Женя | 1       |

Неизвестное слово:  
Иннокентий



# Применение

- Оценка вхождений в документ части поискового запроса (пассажи)
- Оценка части запроса на вхождение известных пассажей

**[где приобрести бесплатно] [георгиевская ленточка]**

|  |                                       |
|--|---------------------------------------|
| $p(\text{где} \mid \langle s \rangle)$               | $= [2\text{gram}] 0.0276706$          |
| $p(\text{приобрести} \mid \text{где} \dots)$         | $= [3\text{gram}] 0.000907595$        |
| $p(\text{бесплатно} \mid \text{приобрести} \dots)$   | $= [3\text{gram}] 0.00042164$         |
| $p(\text{георгиевская} \mid \text{бесплатно} \dots)$ | $= [1\text{gram}] 6.14581\text{e-}07$ |
| $p(\text{ленточка} \mid \text{георгиевская} \dots)$  | $= [2\text{gram}] 0.568604$           |
| $p(\langle /s \rangle \mid \text{ленточка} \dots)$   | $= [3\text{gram}] 0.20317$            |



# Что дальше?

- Улучшаем сглаживание - следующая лекция
- Марковские цепи и Скрытые Марковские Модели.

# МАРКОВСКИЕ МОДЕЛИ

# Цепи Маркова



*Doudou sleeping*



*Doudou eating*



*Doudou training*



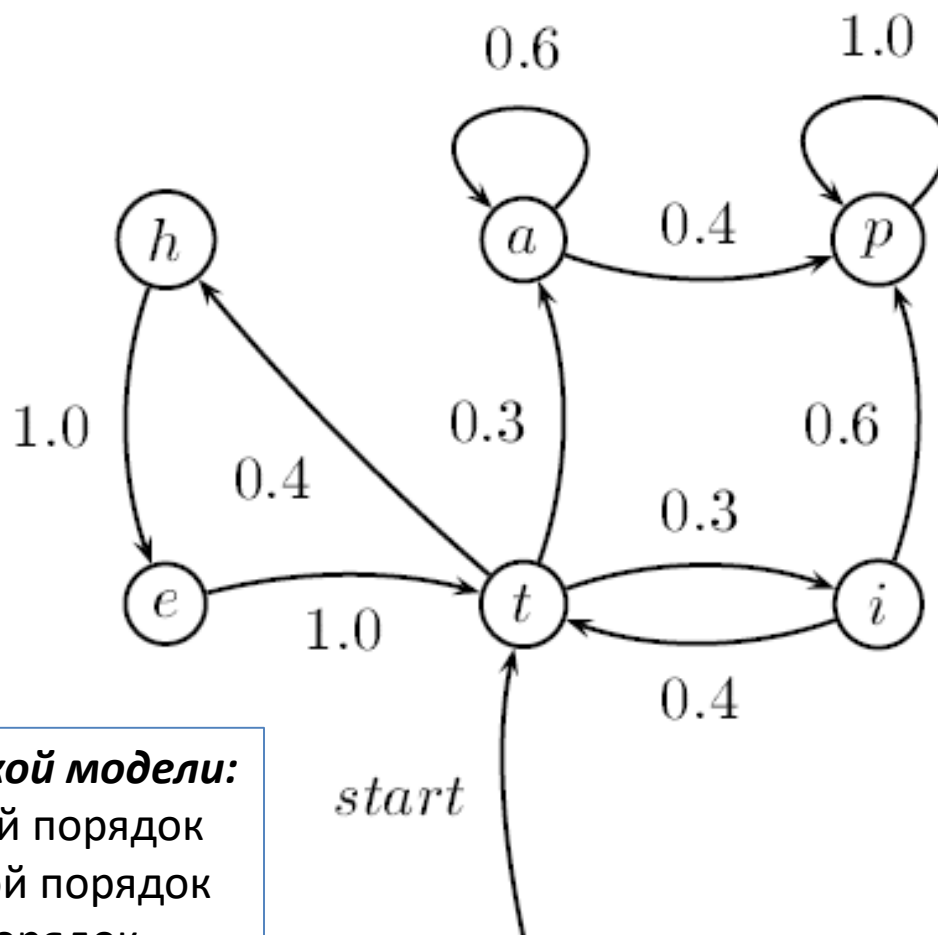
# Марковская модель

- $X = (X_1, \dots, X_T)$  - последовательность случайных величин
- $S = \{s_1, \dots, s_n\}$  - множество состояний этой случайной величины

# Свойства

- Ограниченный горизонт
  - $P(X_{t+1} = s_k \mid X_1, \dots, X_t) = P(X_{t+1} = s_k \mid X_t)$
- Стационарность. Временная инвариантность
  - $P(X_{t+1} = s_k \mid X_t) = P(X_2 = s_k \mid X_1)$
- $A$  – стохастическая матрица переходов
  - $a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i) ; a_{ij} > 0, \forall i, j$  и  $\sum_{j=1}^N a_{ij} = 1$
  - $\pi_i = P(X_1 = s_i)$  - начальное состояние  $\sum_i \pi_i = 1$

# Граф модели



**Порядок Марковской модели:**

Биграммы – первый порядок

Триграммы – второй порядок

$n$  – gram -  $(n-1)$  - порядок

# Примеры

- Call center – перенаправление звонков операторам
- Случайное блуждание по Интернету (PR)
- Модель кликов запрос документ.
- Последовательности слов в тексте

# Последовательность состояний

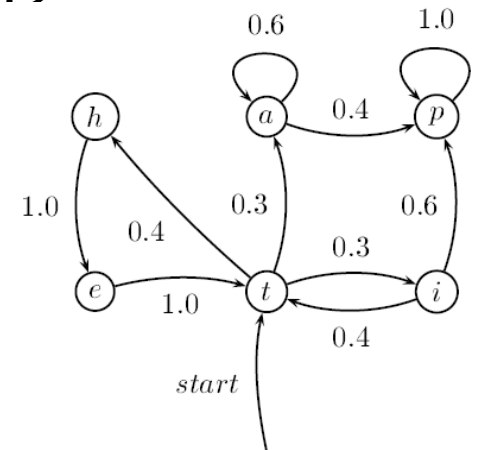
Вероятность последовательности:

- $$P(X_1, \dots, X_T) =$$

$$P(X_1)P(X_2|X_1) \dots P(X_T|X_1, \dots, X_{T-1}) =$$

$$P(X_1)P(X_2|X_1) \dots P(X_T|X_{T-1}) =$$

$$\pi_{x_1} \prod_{t=1}^T a_{X_t X_{t+1}}$$



- $$P(t, i, p) = P(t)P(X_2 = i | X_1 =$$

$$t)P(X_3 = p | X_2 = i) = 1.0 \times 0.3 \times 0.6 = 0.18$$



# Подсчитаем вероятность

пример из тетрадки

*“Президент США Барак Обама решил сняться в телепередаче с Беар Гриллсом.”*

| probs          |   |
|----------------|---|
| 0.002227       | $p_i$ (президент)                               |
| 0.034506       | $p_s(\text{президент} \rightarrow \text{сша})$  |
| 0.022638       | $p_s(\text{сша} \rightarrow \text{барак})$      |
| 0.931526       | $p_s(\text{барак} \rightarrow \text{обама})$    |
| 0.002606       | $p_s(\text{обама} \rightarrow \text{решить})$   |
| 0.000671       | $p_s(\text{решить} \rightarrow \text{сняться})$ |
| 0.451685       | $p_s(\text{сняться} \rightarrow \text{в})$      |
| 0.000018       | $p_s(\text{в} \rightarrow \text{телепередача})$ |
| 0.014085       | $p_s(\text{телепередача} \rightarrow \text{с})$ |
| 0.000062       | $p_s(\text{с} \rightarrow \text{беар})$         |
| 0.316456       | $p_s(\text{беар} \rightarrow \text{гриллсом})$  |
| $6.518518e-24$ | $p(X_1, \dots, X_n)$                            |
| -53.387395     | $\log(p(X_1, \dots, X_n))$                      |

# Подсчитаем вероятность

пример из тетрадки

*“Президент Китая Барак Обама решил сняться в телепередаче с Беар Гриллсом.”*

| probs          |   |
|----------------|---|
| 0.002227       | $p_i$ (президент)                                 |
| 0.000004       | $p\_s(\text{президент} \rightarrow \text{китай})$ |
| 0.000004       | $p\_s(\text{китай} \rightarrow \text{барак})$     |
| 0.931526       | $p\_s(\text{барак} \rightarrow \text{обама})$     |
| 0.002606       | $p\_s(\text{обама} \rightarrow \text{решить})$    |
| 0.000671       | $p\_s(\text{решить} \rightarrow \text{сняться})$  |
| 0.451685       | $p\_s(\text{сняться} \rightarrow \text{в})$       |
| 0.000018       | $p\_s(\text{в} \rightarrow \text{телепередача})$  |
| 0.014085       | $p\_s(\text{телепередача} \rightarrow \text{с})$  |
| 0.000062       | $p\_s(\text{с} \rightarrow \text{беар})$          |
| 0.316456       | $p\_s(\text{беар} \rightarrow \text{гриллсом})$   |
| $1.591735e-31$ | $p(X_1, \dots, X_n)$                              |
| -70.915313     | $\log(p(X_1, \dots, X_n))$                        |

# Сравним вероятности фраз

*P1(Президент США Барак Обама решил сняться в телепередаче с Беар Гриллсом.)*

и

*P2(Президент Китая Барак Обама решил сняться в телепередаче с Беар Гриллсом)*

$$P1(6.518518e-24) > P2(1.591735e-31)$$

*Первая фраза больше подходит под нашу модель*

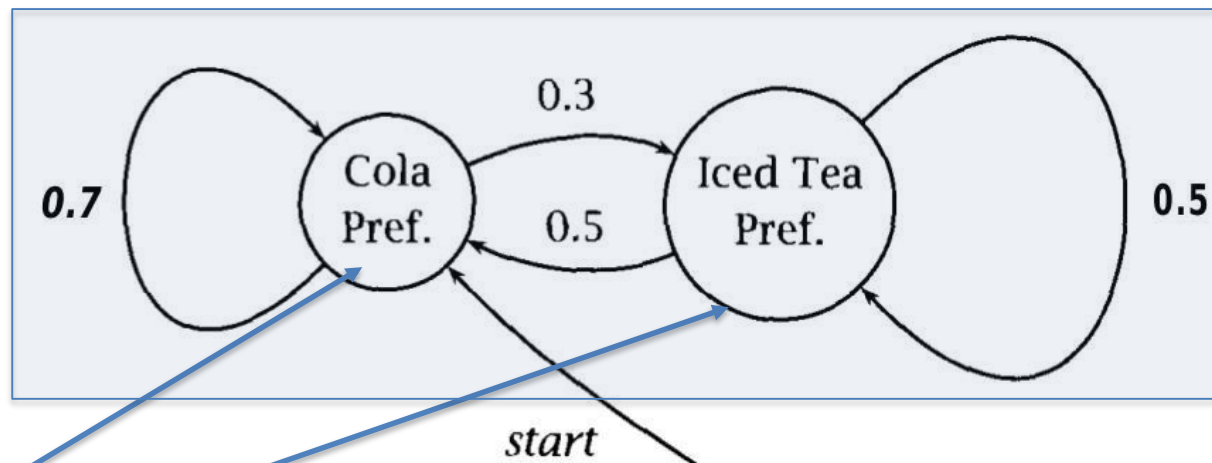
# Примеры для текстов

- Оцениваем авторство человека. Каждый автор имеет свой стиль
- Оценка релевантности документа и запроса
- Классификация источников новостей (ДЗ)

# Автомат прохладительных напитков



WWW.3DVSEM.COM



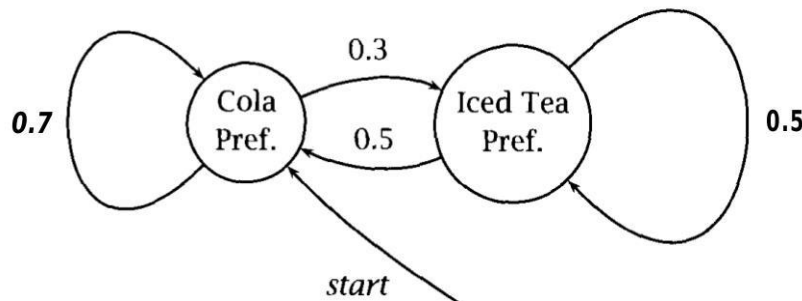
| Два<br>состояния | Кратко |
|------------------|--------|
| Cola Prefect     | CP     |
| Ice Tea Pref     | IP     |

| Состояния | Разливаемые напитки |                    |                   |
|-----------|---------------------|--------------------|-------------------|
|           | cola                | Ice tea<br>(ice_t) | Lemonade<br>(lem) |
| CP        | 0.6                 | 0.1                | 0.3               |
| IP        | 0.1                 | 0.7                | 0.2               |

# Скрытая Марковская модель

- Задача определить вероятность выпуска “символа” не имея информации о состоянии модели.
- $P(O_n = k | X_n = s_i, X_{n+1} = s_j) = b_{ijk}$

# Пример



| Состояния | Разливаемые напитки |                 |                |
|-----------|---------------------|-----------------|----------------|
|           | cola                | Ice tea (ice_t) | Lemonade (lem) |
| CP        | 0.6                 | 0.1             | 0.3            |
| IP        | 0.1                 | 0.7             | 0.2            |

Какова вероятность увидеть последовательность {lem, ice\_t} если машина стартует в CP состоянии?

$$0.3 \times 0.3 \times 0.7 + 0.3 \times 0.7 \times 0.1 = 0.084$$

# Основные термины

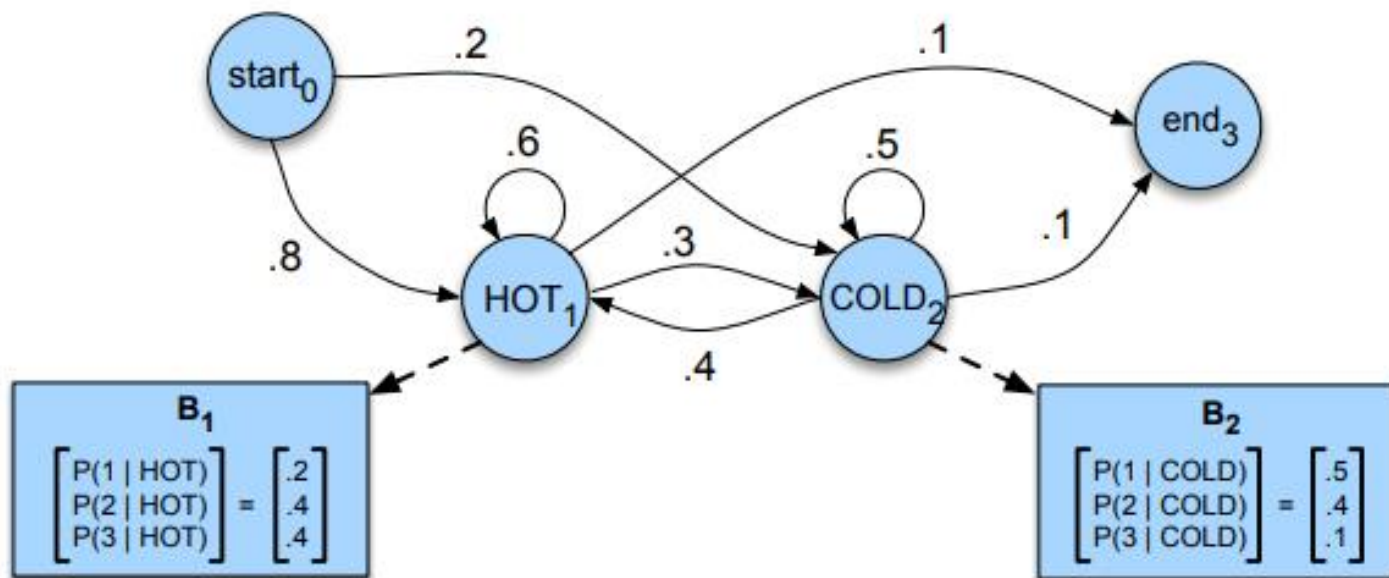
- $S = \{s_1, \dots, s_N\} = \{1 \dots N\}$  – состояния модели
- $K = \{k_1, \dots, M\}$  - алфавит
- $\Pi = \{\pi_i\}, i \in S$  – вероятности начальных состояний
- $A = \{a_{ij}\}, i, j \in S$  - вероятности перехода
- $B = \{b_{jk}\}, j \in S, k \in K$  – вероятность “символа”
- $X = (X_1 \dots X_{T+1})$  – последовательность состояний
- $O = (o_1 \dots o_T)$  – выходная последовательность
- Аналогично Марковской модели:
  - $P(o_i | s_1, \dots s_i, \dots s_T, o_1, \dots o_i, \dots o_T) = P(o_i | s_i)$



# Три задачи НММ

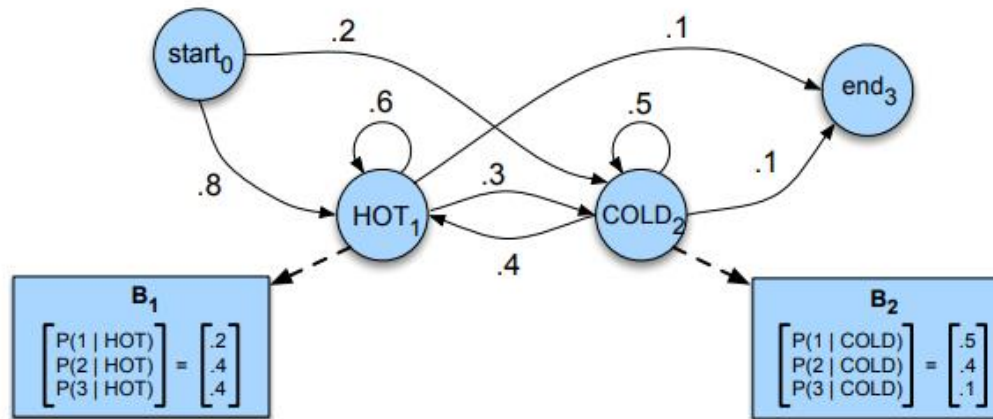
1. Дано: модель  $\mu(A, B, \Pi)$  и наблюдения  $O$ .
  - Оценить насколько наши наблюдения под модель, т.е. оценить вероятность  $P(O|\mu)$
2. Дано: модель  $\mu$  и наблюдения  $O$ .
  - Выбрать последовательность  $(X_1 \dots X_{T+1})$ , которая лучше описывает наши наблюдения
3. Дано: последовательность наблюдений  $O$  и словарь состояний  $K$ 
  - Найти модель, лучше описывающую наши наблюдения  $\mu(A, B, \Pi)$

# Погода по мороженому



# Задача 1

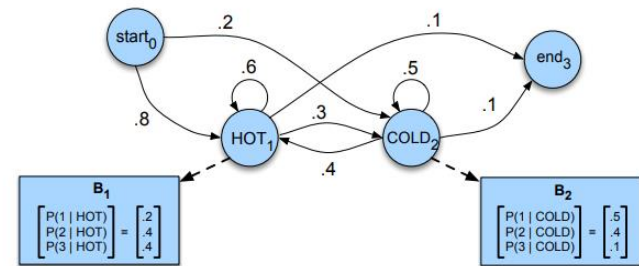
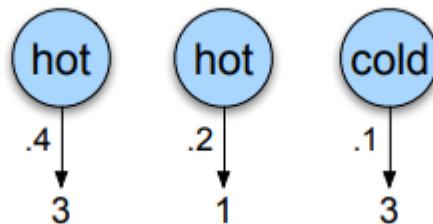
- Оценить вероятность появления последовательности 3-1-3



- Если мы наблюдаем последовательность *hot – hot – cold* – то сделать это просто

# Задача 1

- $P(O|S) = \prod_{i=1}^T P(o_i|s_i)$



$$P(3 \ 1 \ 3 | \text{hot hot cold}) = P(3|\text{hot}) \times P(1|\text{hot}) \times P(3|\text{cold})$$

- Но мы не знаем какая реальная последовательность погоды была

# Задача 1

- Поэтому нам нужна сумма совместных вероятностей событий погоды и последовательностей 3-1-3:
- $P(O, S) = P(O|S)P(S) = \prod_{i=1}^T P(o_i|s_i) \prod_{i=1}^T P(s_i|s_{i-1})$  - совместная вероятность

$$P(3 \ 1 \ 3, \text{hot hot cold}) = P(\text{hot}|\text{start}) \times P(\text{hot}|\text{hot}) \times P(\text{cold}|\text{hot}) \\ \times P(3|\text{hot}) \times P(1|\text{hot}) \times P(3|\text{cold})$$

- Теперь посчитаем вероятность наших наблюдений
- $P(O) = \sum_S P(O, S) = \sum_S P(O|S)P(S)$

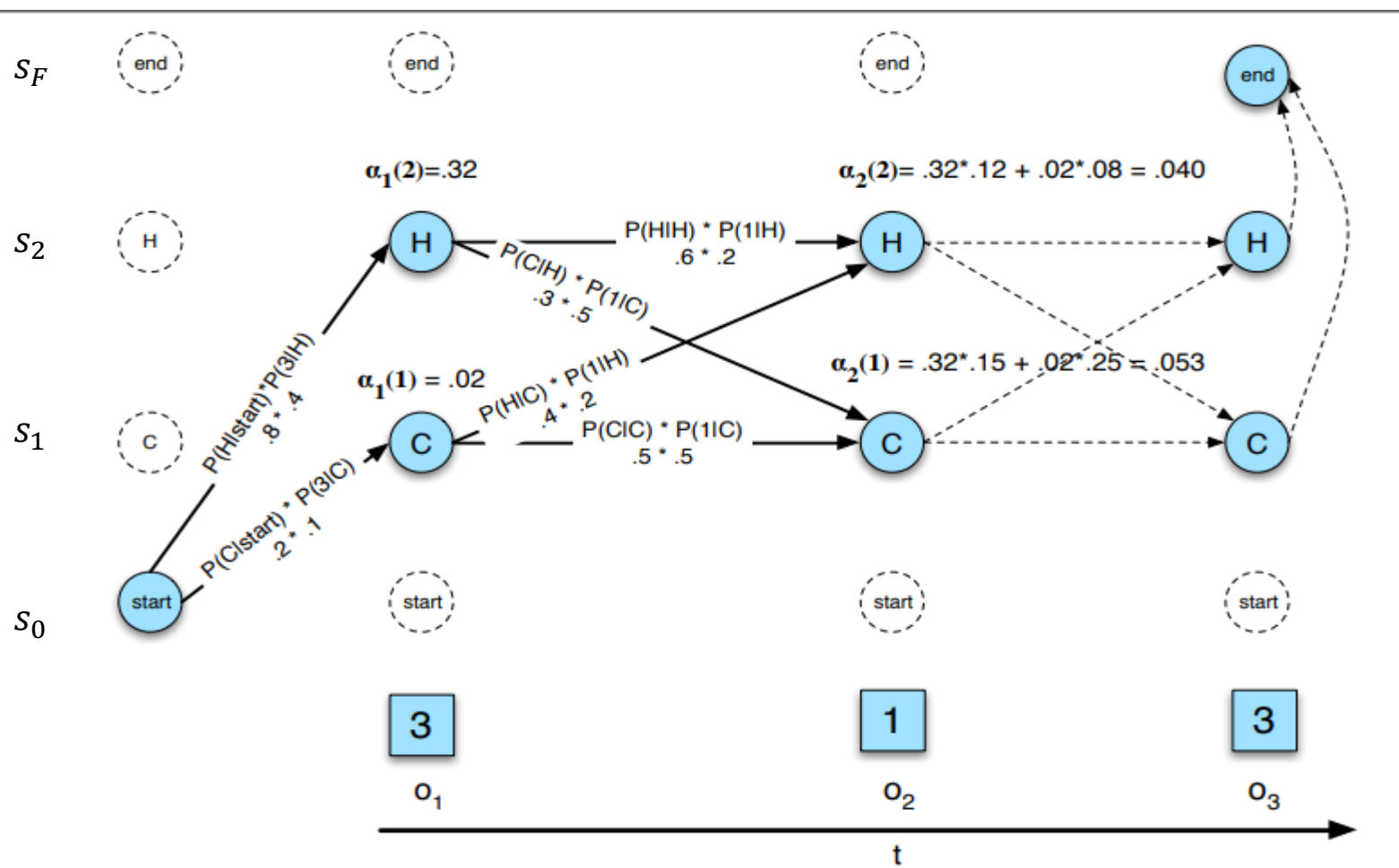
# Задача 1

- Для нашей последовательности 3-1-3

$$P(3\ 1\ 3) = P(3\ 1\ 3, \text{cold cold cold}) + P(3\ 1\ 3, \text{cold cold hot}) + P(3\ 1\ 3, \text{hot hot cold}) + \dots$$

- Сложность  $O(N^T)$ :
  - $N$  – количество скрытых состояний
  - $T$  – количество наблюдений
- Выход – динамическое программирование и алгоритм прямого прохода
  - Сложность -  $O(N^2T)$ :

# Задача 1



# Задача 1

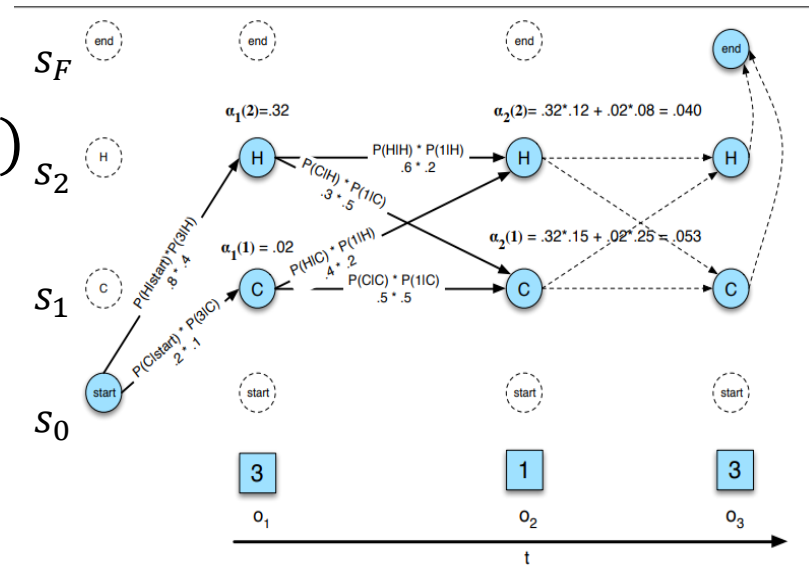
- $\alpha_t(j)$  - вероятность нахождения в узле  $j$  – после наблюдения первых  $t$  наблюдений задаваемых моделью  $\mu$

Каждый элемент сетки выражает следующую вероятность:

$$\alpha_t(j) = P(o_1 o_2 \dots o_t, s_t = j | \mu)$$

или

$$\alpha_t(j) = \sum_{i=1}^N \alpha_{t-1}(i) a_{ij} b_{jo_t}$$





# Прямой проход алгоритм

- Инициализация:

$$- \alpha_t(j) = \pi_j b_{j1}, 1 \leq i \leq N$$

- Рекурсивно считаем:

$$- \alpha_t(j) = \sum_{i=1}^N a_{t-1}(i) a_{ij} b_{j o_t}, 1 \leq t \leq T, 1 \leq j \leq N$$

- Окончание

$$- P(O|\mu) = \sum_{i=1}^N \alpha_T(i)$$

# Пример: выделение адресов

- Идея – выделяем ключевую точку по словарю (город), и оцениваем окрестность как вероятность  $P(O|\mu)$

|          |              |
|----------|--------------|
| <i>C</i> | страна       |
| <i>S</i> | улица        |
| <i>D</i> | район города |
| <i>H</i> | номер дома   |
| <i>F</i> | квартира     |
| <i>T</i> | город        |
| <i>O</i> | разделители  |
| ...      |              |

|  |        |      |             |      |      |              |   |      |      |
|--|--------|------|-------------|------|------|--------------|---|------|------|
| <div> <math>P(O \mu)</math> <math>P(O \mu)</math> </div> |        |      |             |      |      |              |   |      |      |
| “Выпуск” символов <i>O</i> – наблюдения                  |        |      |             |      |      |              |   |      |      |
|  | I      | O    | T           | O    | ms   | S            |   | O    | H    |
|  | 654007 | ,    | Новокузнецк | ,    | ул.  | Орджоникидзе | , |      | 36   |
| <i>V</i>   | v(1)   | v(2) | v(3)        | v(2) | v(4) | v(5)         |   | v(2) | v(6) |
| <i>S</i>   | s(1)   | s(2) | s(3)        | s(2) | s(4) |              |   | s(2) | s(5) |

O

X

Комбинируем оба прохода

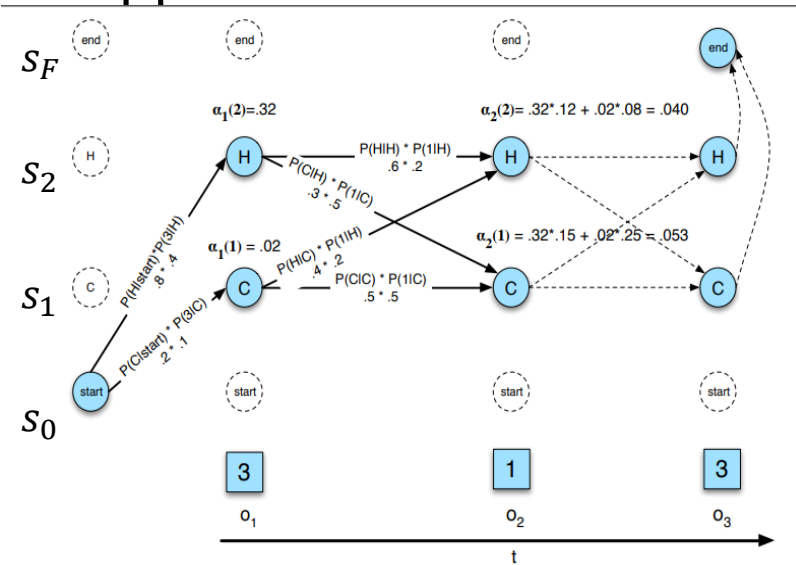
$$P(O|\mu) = \sum_{i=1}^N \alpha_t(i) \beta_t(i)$$

## Задача 2 - Декодирование

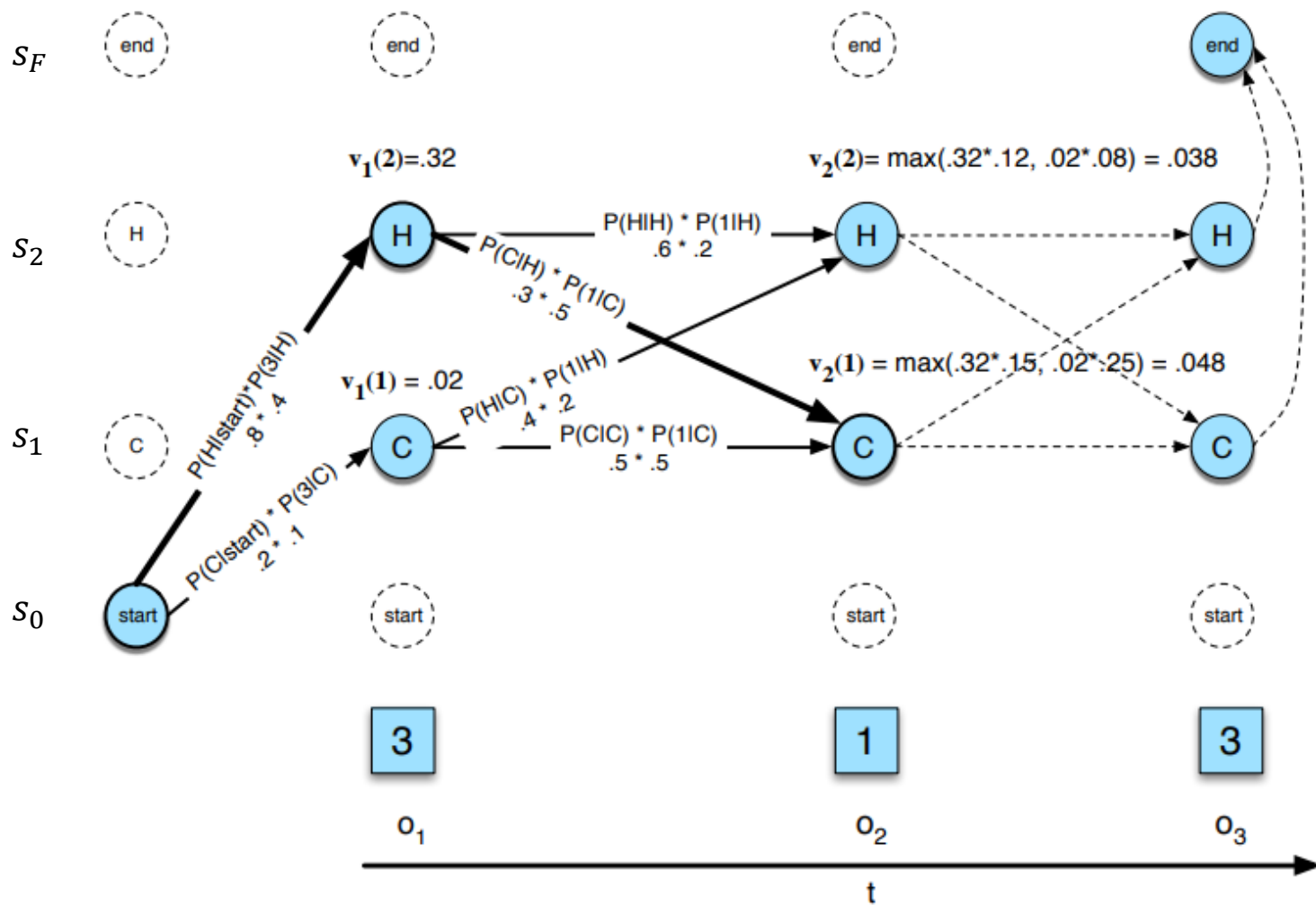
- Дано
  - модель:  $\mu(A, B)$
  - последовательность наблюдений  $O = o_1, o_2, \dots o_T$
- Найти последовательность наиболее вероятных скрытых состояний  $S = s_1, s_2, \dots s_T$
- В случае погоды по мороженному нужно определить какая была погода по последовательности 3-1-3

## Задача 2

- Наивный подход – запустить прямой проход на перебор состояний:
  - hhh, hhc, hch ...
  - Затем выбрать лучшую последовательность
- Выход динамическое программирование и алгоритм Витерби.



# Задача 2 - Алгоритм Витерби



## Задача 2

- $v_t(j)$  - показывает вероятность того, что модель находится в наиболее вероятном состоянии  $j$  после наблюдения  $t$

Каждый элемент сетки выражает следующую вероятность:

$$v_t(j) = \max_{s_1, s_2, \dots, s_{t-1}} P(s_1, s_2, \dots, s_{t-1}, o_1, o_2, \dots, o_t, \quad s_t = j | \mu)$$

или выразив  $v_t(j)$  через предыдущее состояние:

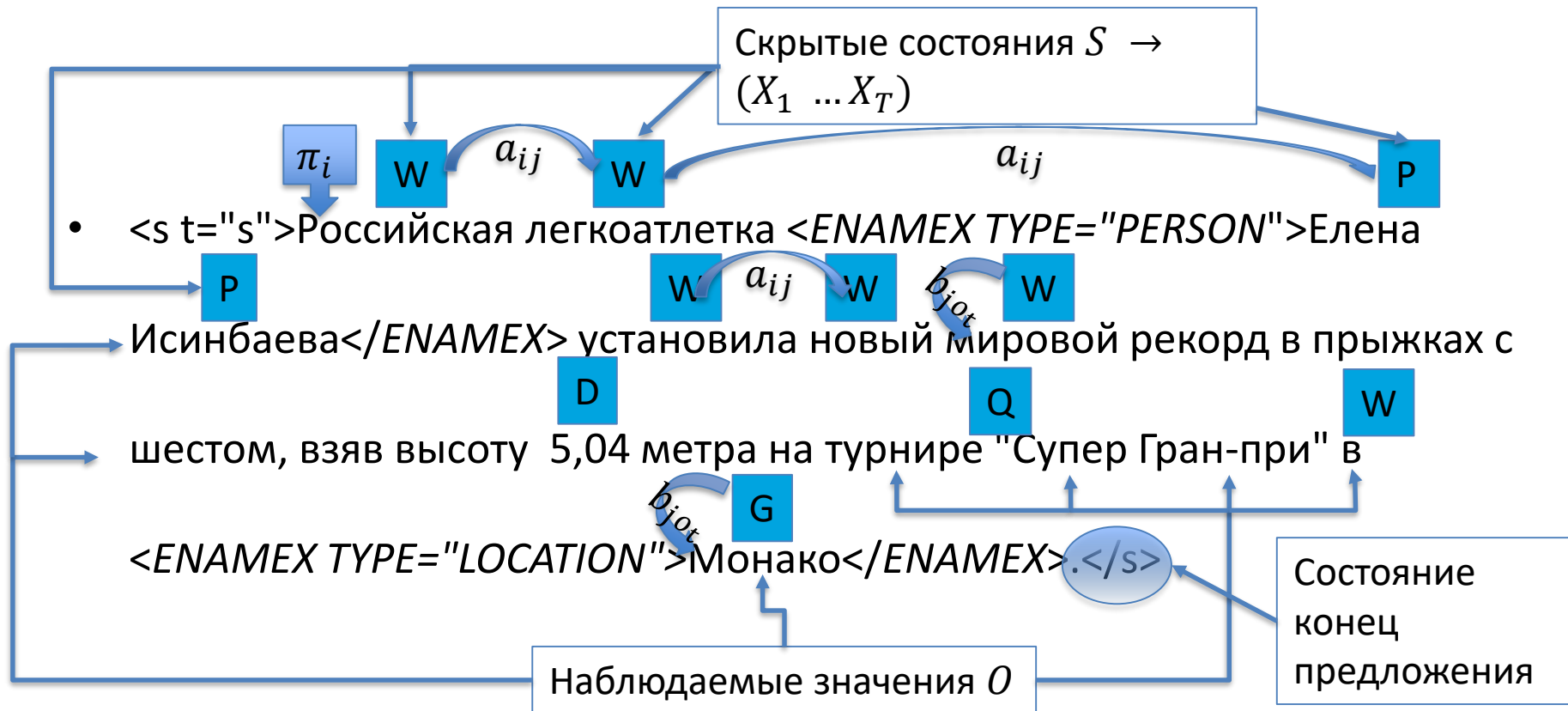
$$v_t(j) = \max_{i=1-N} v_{t-1}(i) a_{ij} b_{jo_t}$$

# Алгоритм Витерби

- Инициализируем:
  - $v_1(j) = \pi_j b_{j1}, 1 \leq j \leq N; \psi_1(j) = []$
- Рекурсивно считаем:
  - $v_t(j) = \max_{1 \leq i \leq N} [v_{t-1}(i) a_{ij}] b_{j o_t}, 1 \leq j \leq N$
  - Сохраняем состояние:  $\psi_t(j) = \operatorname{argmax}_i [\delta_{t-1}(i) a_{ij}] b_{j o_t}, 1 \leq j \leq N$
- Заканчиваем, и считываем, что получилось:
  - $\hat{X}_T = \operatorname{argmax}_i v_T(i)$  -финал
  - $\hat{X}_t = \psi_{t+1}(\hat{X}_{t+1})$  - тут путь, который нам нужен
  - $P(\hat{X}) = \max_i v_T(i)$
- Возвращаем список лучших состояний.

Обратный указатель

# Пример: NER



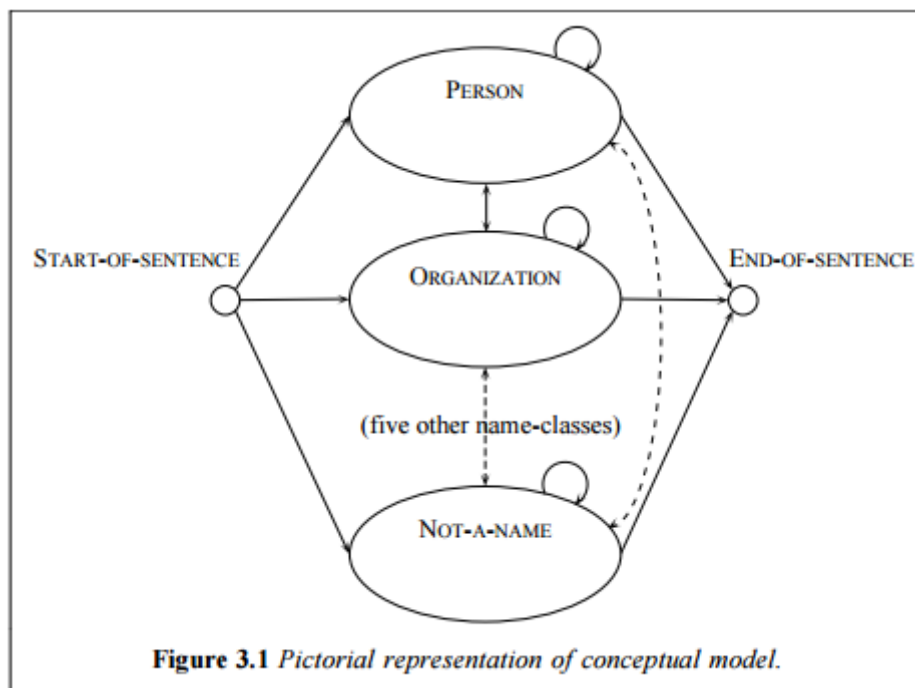
Используем алгоритм Витерби для декодирования скрытой последовательности по нашим наблюдениям

[Daniel M. Bikel 1997, Nymble: a High-Performance Learning Name-finder](#)



# Тегинг

| Word Feature           | Example Text                  |
|------------------------|-------------------------------|
| twoDigitNum            | 90                            |
| fourDigitNum           | 1990                          |
| containsDigitAndAlpha  | A8956-67                      |
| containsDigitAndDash   | 09-96                         |
| containsDigitAndSlash  | 11/9/89                       |
| containsDigitAndComma  | 23,000.00                     |
| containsDigitAndPeriod | 1.00                          |
| otherNum               | 456789                        |
| allCaps                | BBN                           |
| capPeriod              | M.                            |
| firstWord              | <i>first word of sentence</i> |
| initCap                | Sally                         |
| lowerCase              | can                           |
| other                  | ,                             |



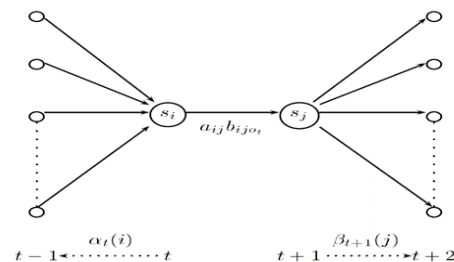
## Задача 3 - Обучение

- Дано
  - Последовательность наблюдений  $O = O_1, O_2, \dots, O_T$
  - Словарь состояний  $S$
- Ищем оптимальные значения модельных параметров:  $\mu(A, B, \pi)$
- Используем итеративный алгоритм “Вперед Назад” или “Baum-Welch” или EM

# Задача 3

- Рассмотрим Марковскую модель:
  - Состояния известны
  - Примем, что вероятность  $b = 1$
  - Тогда вероятность перехода из состояния  $i$  в  $j$

$$a_{ij} = \frac{C(i \rightarrow j)}{\sum_{s \in Q} C(i \rightarrow s)}$$



- *В НММ- мы не можем знать об этих состояниях*

## Задача 3 – Идея 1

- Мы можем оценить вероятность перехода итеративно
  - Оцениваем вероятность перехода и наблюдений
  - Затем используем эти оценки для улучшения следующих оценок

## Задача 3 – Идея 2

- Мы можем использовать **прямые вероятности**  $\alpha_t(j)$
- Так же нам потребуется научиться рассчитывать обратные вероятность. Это вероятность увидеть наши наблюдения с момента времени  $t+1$  до  $T$

## Задача 3. Обратный проход

- Принимаем, что:
  - $\beta_t(i) = P(o_t \dots o_T, |s_t = i, \mu)$
- Инициализация:
  - $\beta_t(i) = 1, 1 \leq i \leq N$
- Рекурсивно:
  - $\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) a_{ij} b_{j o_{t+1}}, 1 \leq t \leq T, 1 \leq i \leq N$
- $P(O|\mu) = \sum_{i=1}^N \pi_i b_{j o_1} \beta_1(i)$

## Задача 3

- Теперь нам нужно оцениваем вероятность перехода из состояния  $i$  в  $j$

$$- \hat{a}_{ij} = \frac{\text{ожидаемое количество переходов из } i \text{ в } j}{\text{ожидаемое количество переходов из } i}$$

## Задача 3

- Введем вероятность  $\varepsilon_t(i, j)$  – нахождения системы в состоянии  $i$  в момент  $t$  и в состоянии  $j$  в момент  $t+1$   
–  $\varepsilon_t(i, j) = P(s_t = i, s_{t+1} = j \mid O, \mu)$

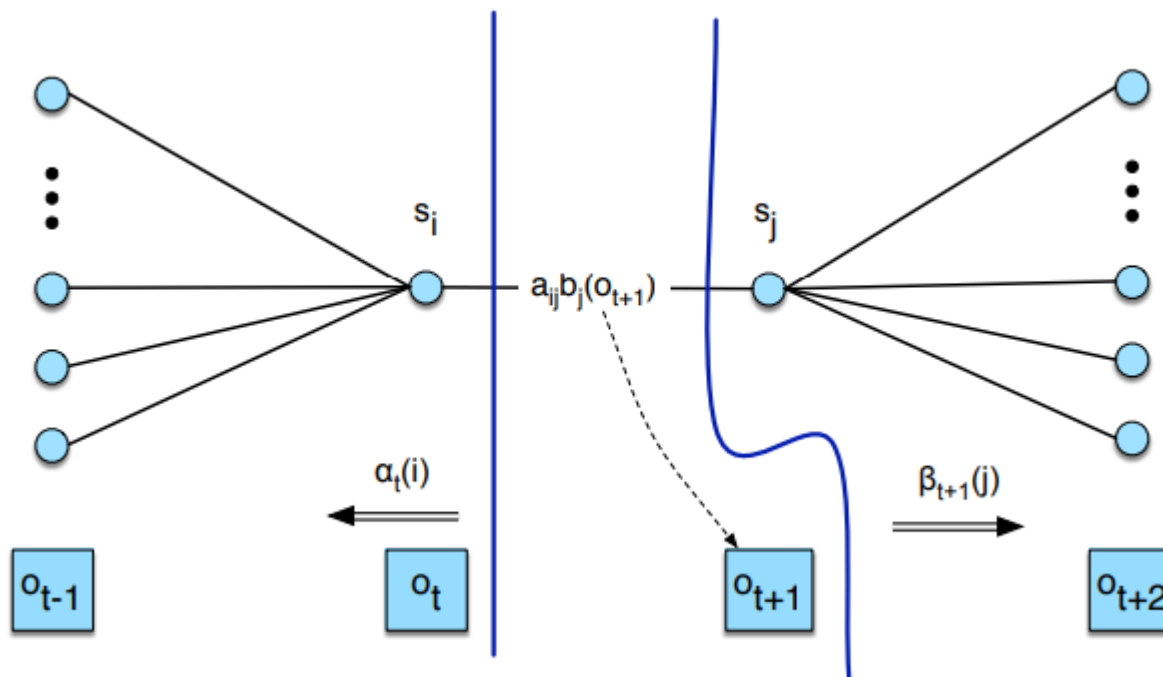
Введем доп вероятность:

$$nq\varepsilon_t(i, j) = P(s_t = i, s_{t+1} = j, O \mid \mu)$$



# Задача 3 - $nq\varepsilon_t(i, j)$

$$nq\varepsilon_t(i, j) = \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$



## Задача 3 - $\varepsilon_t(i, j)$

Используем свойство:

$$P(X|Y, Z) = \frac{P(X, Y|Z)}{P(Y|Z)}$$

Имеем:

$$P(O|\mu) = \sum_{j=1}^N a_t(j) \beta_t(j)$$

Тогда

$$\varepsilon_t(i, j) = \frac{\alpha_t(i) a_{ij} b_{ijo_t} \beta_{t+1}(j)}{\sum_{m=1}^N \alpha_t(m) \beta_t(m)}$$

## Задача 3 - $\hat{a}_{ij}$

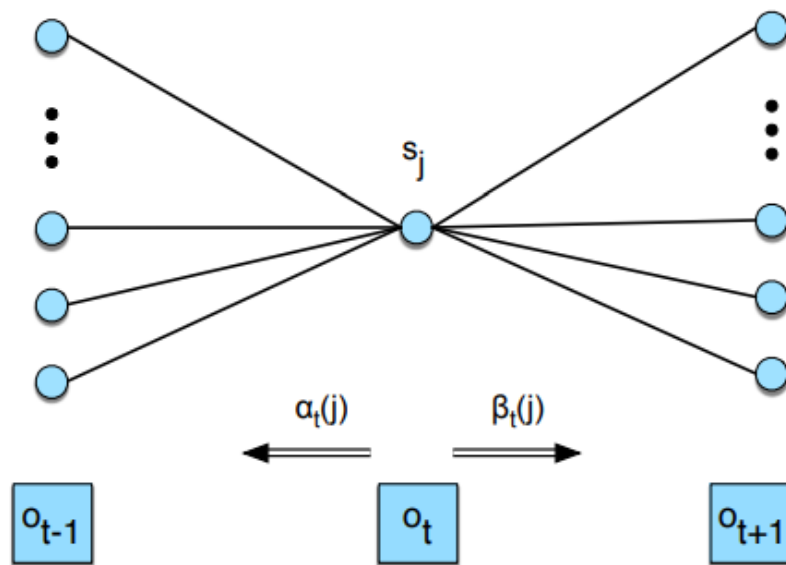
$$\hat{a}_{ij} = \frac{\sum_{t=1}^T \varepsilon_t(i, j)}{\sum_{t=1}^{T-1} \sum_k^N \varepsilon_t(i, k)}$$

# Задача 3 - $\hat{b}_{jk}$

Вероятность выпуска символа  $k$  в состоянии  $j$

$$\hat{b}_{jk} = \frac{\text{ожидаемое количество появлений в } j \text{ с наблюдением } k}{\text{ожидаемое количество появлений в } j}$$

$$\begin{aligned} \gamma_t(i) &= P(s_t = i | O, \mu) \\ &= \frac{P(X_t = i, O | \mu)}{P(O | \mu)} \\ &= \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)} \end{aligned}$$



## Задача 3 - $\hat{b}_{jk}$

Вероятность выпуска символа  $k$  в состоянии  $j$

$$\hat{b}_{jk} = \frac{\text{ожидаемое количество появлений в } j \text{ с наблюдением } k}{\text{ожидаемое количество появлений в } j}$$

$$\gamma_t(i) = P(s_t = i | O, \mu) = \frac{P(X_t = i, O | \mu)}{P(O | \mu)} = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)}$$

$$\hat{b}_{jk} = \frac{\sum_{\{t: o_t=k, 1 \leq t \leq T\}} \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)}$$

# Алгоритм “Вперед- назад”

- Начинаем с некоторой модели  $\mu$ :
  - строим по нашим данным
  - инициализируем случайно

Шаги:

- E: Скармливаем модели наши наблюдения  $O$  и оцениваем ожидание модельных параметров
- M: Затем переоцениваем параметры модели:
  - $\hat{\pi}_i$  = ожидаемая частота в состоянии  $i$  в момент  $t = 1$ ;
  - $\cdot = \gamma_1(i)$
  - $\hat{a}_{ij} = \frac{\text{ожидаемое количество переходов из } i \text{ в } j}{\text{ожидаемое количество переходов из } i} = \frac{\sum_{t=1}^T \varepsilon_t(i,j)}{\sum_{t=1}^T \gamma_i(t)}$
  - $\hat{b}_{jk} = \frac{\text{ожидаемое количество появлений в } j \text{ с наблюдением } k}{\text{ожидаемое количество появлений в } j} = \frac{\sum_{\{t: o_t=k, 1 \leq t \leq T\}} \gamma_t(j)}{\sum_{t=1}^T \gamma_t(i)}$
- Таким образом:  $\mu(A, B, \Pi) \Rightarrow \hat{\mu} = (\hat{A} \hat{B} \hat{\Pi})$ ;  $P(O|\hat{\mu}) \geq P(O|\mu)$

# Вопросы и ДЗ