

Статистическая обработка текстов

Коллокации, N-граммы, НММ

Андрей Кривой
Руководитель группы ранжирования
Поиск@Mail.Ru

Москва 2018

План

- Лингвистика
 - Коллокации. Методы их нахождения
 - N-граммы. Языковые модели
- Марковские модели
 - Марковские цепи
 - Скрытые Марковские модели (НММ)
- Домашняя работа

Коллокация – это 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

- Примеры:
 - *идет дождь*
 - *высокая температура*
 - *кот в мешке*
 - *strong tea, крепкий чай*
- Коллокация, как правило, состоит из 2-х слов (компонент):
 - свободное: *дождь*
 - несвободное, или зависимое: *идет*

Признаки коллокации

- Некомпозиционность
 - Смысл коллокации не является композицией смысла частей: **высокая температура**
- Незаменяемость
 - Нельзя заменять зависимое слово на другое подходящее по смыслу: **крепкий чай != прочный чай**

Признаки коллокации

- Порядок слов в коллокации важен!
 - *рост влияния* — это коллокация, а *влияния рост* — уже нет
- Между образующими коллокацию словами могут быть разрывы:
 - *идет* осенний *дождь*
 - *идет* мелкий осенний *дождь*

Коллокации vs. Идиомы

Фразеологизм (идиома) — устойчивое сочетание слов, значение которого **не** определяется значением входящих в него слов, взятых по отдельности:

- *делать из мухи слона*
- *купить коша в мешке*
- *kick the bucket*

Идиому можно считать коллокацией с максимальной некомпозиционностью!

Коллокации vs. Имена

- К коллокациям часто причисляют названия
- Топонимы:
 - *Нижний Новгород, Черное море*
- Имена людей (антропонимы):
 - *Александр Сергеевич Пушкин*
- Другие совместно употребляемые наименования:
 - *Университет Информационных Технологий, Механики и Оптики*

Коллокации vs. Ассоциации

- Иногда нужно найти слова, которые часто встречаются в одних и тех же контекстах (ассоциации, co-occurences):
 - *самолет и аэропорт*
 - *весна и цветы*
- В тех случаях, когда такие пары слов **не** являются синтаксически связанными, мы **не** будем считать их коллокациями

Коллокации – и все-таки, что это?

- Сложно дать однозначное определение коллокации, оно зависит от контекста и от наших целей
- В последние годы во главу угла ставится частота совместной встречаемости, т.е. коллокации могут быть определены прежде всего как **статистически устойчивые словосочетания**

Применение

- Лингвистика (лексикография, корпусные исследования, ...)
- Генерация текстов
- Машинный перевод
- Информационный Поиск (information retrieval)

Коллокации и машинный перевод

- При переводе обойтись без знания о коллокациях невозможно!
 - strong tea => **сильный чай**
 - strong tea => **крепкий чай**
- Особенно тяжело переводить идиомы
 - льет как из ведра ⇔ it rains cats and dogs

Коллокации в поиске

- Люди подают запросы в свободной форме
 - ситуация в россии **трезвая оценка**
- Коллокации не должны разрываться в найденных документах:
 - Федеральный проект **Трезвая Россия**
- Найденные в документах коллокации могут быть использованы в качестве признаков для машинного обучения

Коллокации в поиске

- Мы должны учитывать коллокации на этапе расширения поискового запроса синонимами
- *Крепкий – сильный, устойчивый, прочный, твердый, здоровый, надёжный;*
- *крепкий узел*
=> (крепкий | | прочный) && узел
- *крепкий чай*
=> крепкий && чай

Выделение коллокаций

- Как же все-таки искать коллокации в тексте?

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
 - N – существительное (noun)
 - A – прилагательное (adjective)
- Паттерны:
 - AN – учебный год
 - NN – пресс служба
 - AAN - дискретная случайная величина
 - ANN - эмпирическая функция распределения

	Частота
пресс служба	129071
тот число	69728
уголовный дело	57108
риа новость	54353
миллион рубль	52868
настоящее время	44442
миллиард рубль	44368
прошое 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 неразрывных коллокаций, но...
- ‘стучаться в дверь’
 - она стучалась в дверь
 - они стучались в эту деревянную дверь
 - он стучался в металлическую дверь

Среднее и дисперсия

- Используем окно в 3 слова

Экономика балансирует на грани коллапса

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

Среднее и дисперсия

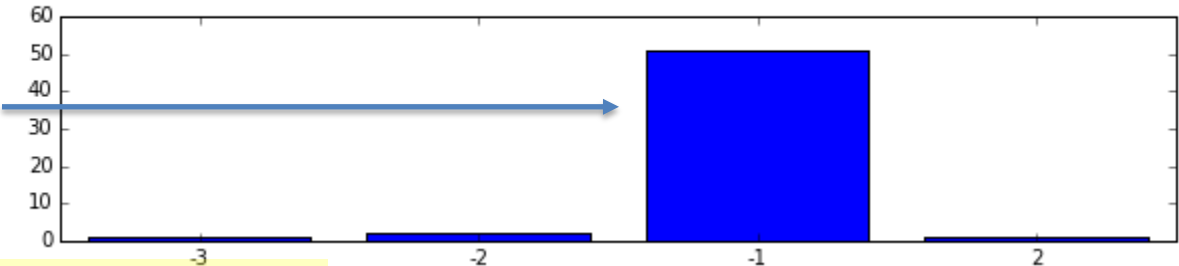
- Пример: 'стучаться ... дверь'
- $\mu = (2+4+3) / 3 = 3$
- $$\sigma^2 = \frac{\sum_{i=1}^n (d_i - \mu)^2}{n-1}$$
$$= \frac{1}{2} (2 - 3)^2 + (4 - 3)^2 + (3 - 3)^2$$
$$= 1$$
- Небольшая дисперсия говорит, что два слова встречаются приблизительно на одном расстоянии

Среднее и дисперсия

Маленькие дисперсия и среднее

сильная волатильность

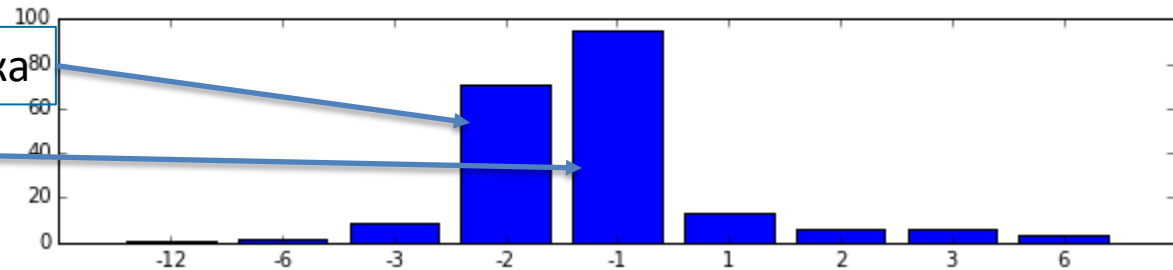
mean: -1.02 std: 0.52/ (волатильность сильный)



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

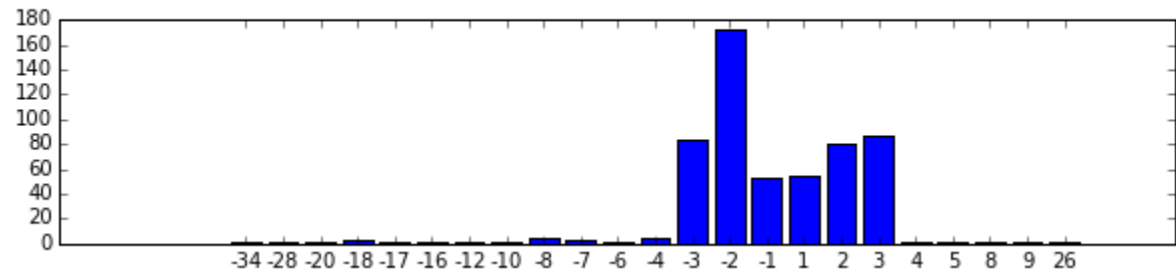
сильная медийная поддержка

сильная поддержка



mean: -0.75 std: 3.84/ (для сильный)

Нет интересных коллокаций



Проверка гипотез

- У предыдущих методов есть один серьезный недостаток: высокая частота и низкая дисперсия могут быть просто случайностью
- Призовем на помощь статистику!
- Сможем ли мы понять, когда два слова встречаются вместе настолько часто, что это уже нельзя назвать случайностью?

Проверка гипотез

- Классическая задача из статистики:
 - Формулируем нулевую гипотезу H_0 :
Между словами нет никакой связи, только случайная встречаемость
 - Вычисляем вероятность p при условии $H_0 = true$
 - Отвергаем H_0 если p ниже некоторого порога (0.05, 0.01, ...)
 - В противном случае принимаем гипотезу H_0

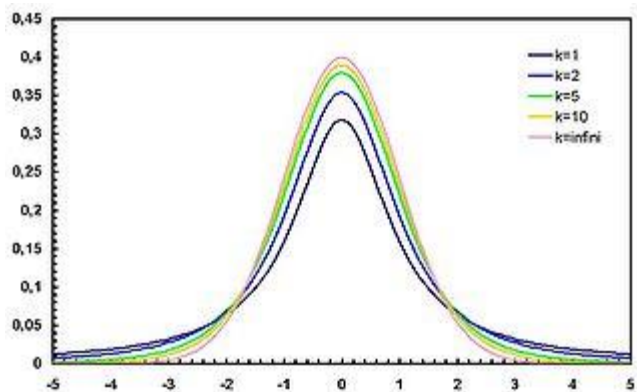
Проверка гипотез

- Как применить проверку гипотез в случае коллокаций?
 - $H_0 = true$ если два слова (w_1, w_2) не образуют коллокацию
 - Предполагаем, что w_1 и w_2 генерируются независимо друг от друга:
$$P(w_1, w_2) = P(w_1)P(w_2)$$
 - Не совсем точная, но простая модель, которая подходит для решения задачи

Распределение Стюдента

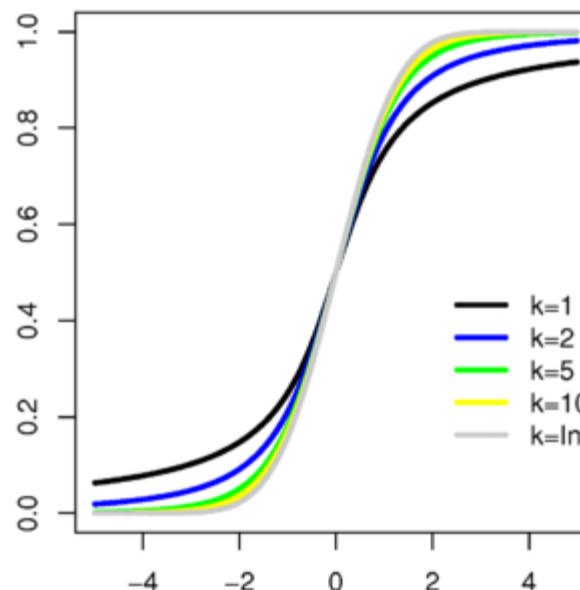
- Пусть X_0, X_1, \dots, X_n – независимые одинаково распределенные случайные величины
- $X_i \sim N(\mu, \sigma^2), i = 0, \dots, n$
- \bar{X}_n – выборочное среднее
- S_n – выборочное стандартное отклонение
- Рассмотрим случайную величину T:
- $T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}}$ распределение Стюдента с (n-1) степенями свободы

Распределение Стьюдента



Плотность вероятности

Для этого распределения
рассчитаны таблицы для
разных степеней свобод.



Функция распределения

t-критерий Стьюдента

- Допущение – примеры отбираются из нормального распределения
- Нулевая гипотеза H_0 : выборка сэмплирована из распределения с мат. ожиданием μ_0
- Оценка среднего: $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$
- Оценка дисперсии: $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$
- Вычисляем величину: $t = \frac{x_n - \mu_0}{s_n / \sqrt{n}}$

t-критерий Стьюдента

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

		Достоверность					
		P	0.05	0.025	0.01	0.005	0.001
		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	4.144
		20	1.725	2.086	2.528	2.845	3.552
	(z)	∞	1.645	1.960	2.326	2.576	3.091

t-критерий Стьюдента

- H_0 гипотеза: средний рост мужчин 158 см
- У нас есть выборка из 200 человек:
 - $\bar{x}_n = 168, s_n^2 = 2600$
 - $t = \frac{169 - 158}{\sqrt{2600/200}} \approx 3,05$
 - при уровне достоверности $\alpha = 0.005$ находим в таблице значение: 2,576.
 - $3,05 > 2,576 \Rightarrow$ с 99% вероятностью отвергаем нулевую гипотезу

t–критерий для коллокаций

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

t-критерий для коллокаций

- Если $H_0 = true \Rightarrow$
 - процесс генерации одной биграммы подчиняется распределению Бернулли с $p = 0.0000017$
 - 1 – если (биграмма == ‘курс рубля’)
 - 0 – если (биграмма != ‘курс рубля’)
 - $\mu_0 = p$ – значения мат. ожидания, которое мы хотим подтвердить либо опровергнуть
 - Оценим: $s_n^2 = p(1 - p) = 0.00000166 \approx p$

t–критерий для коллокаций

- Для биграммы оценим реальное среднее

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

- Проводим t–тест при $\alpha = 0.005$:

$$t = \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}} = 228.58 \gg 2.576$$

- У нас $n \gg 0 \Rightarrow \text{d.f.} \sim \infty$
- Мы можем отвергнуть гипотезу H_0 о независимости слов!

Критерий Пирсона

- Критерий Пирсона (критерий χ^2)
 - Не требует нормального распределения (в отличие от t-критерия)
 - Сравнивает наблюдаемые частоты с ожидаемыми в случае независимости событий
 - Проверяет соответствие наблюдаемой выборки некоторому теоретическому закону распределения

Критерий Пирсона

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

- «Новая компания» -- это коллокация?
- $C(\text{новый}) = 211876$
- $C(\text{компания}) = 233599$
- $C(\text{новая компания}) = 335$
- Всего биграмм: 186419524

Критерий Пирсона

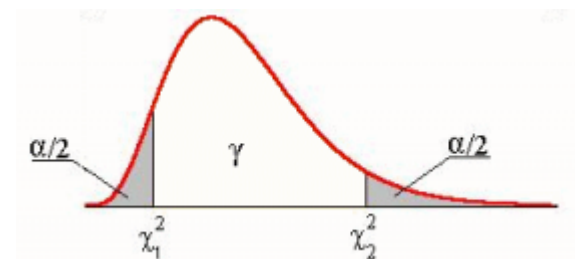
- Суммируем разницу между наблюдаемыми и ожидаемыми значениями

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

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

- X^2 асимптотически сходится к χ^2

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



Критерий Пирсона для коллокаций

		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 может быть отвергнута

Проверка гипотез

- Получается, ‘новая компания’ – это коллокация?
- В естественном языке большинство биграмм – не случайны!
- Мы можем ранжировать биграммы по значениям статистик t и χ^2 и считать коллокациями те, что проходят по какому-то порогу отсеечения

Отношение правдоподобия

- Он же Likelihood Ratio Test (LRT)
- Лучше чем критерии Стьюдента и χ^2 для разреженных данных
- Дает более понятную интерпретацию результата, т.е. во сколько раз одна гипотеза лучше другой

Отношение правдоподобия

- Рассматриваем две гипотезы:
 - H_1 : какой-то из параметров модели $\theta = \theta_1$
 - H_2 : какой-то из параметров модели $\theta = \theta_2$
- Считаем отношение: $\log \lambda = \frac{L(\theta_1; X=x)}{L(\theta_2; X=x)}$
- Можно доказать, что $-2 \log \lambda \sim \chi^2 \Rightarrow$
можно пользоваться таблицами для
проверки гипотезы H_1

Отношение правдоподобия

- Попробуем применить LRT для коллокаций
- Сравниваем между собой две гипотезы:
 - $H_1: P(w_2 | w_1) = p = P(w_2 | \neg w_1)$ – гипотеза о независимости
 - $H_2: P(w_2 | w_1) = p_1 \neq p_2 = P(w_2 | \neg w_1)$
гипотеза о зависимости

Отношение правдоподобия

- c_1, c_2 и c_{12} – частоты слов w_1, w_2 и биграммы $w_1 w_2$
- p – абсолютная вероятность w_2
- p_1 и p_2 – условные вероятности слова w_2 в тех случаях, когда перед ним есть или нет слова w_1
- $p = \frac{c_2}{N}; p_1 = \frac{c_{12}}{c_1}; p_2 = \frac{c_2 - c_{12}}{N - c_1}$
- Предполагаем, что кол-во биграмм в корпусе подчиняется биномиальному распределению:
$$b(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Отношение правдоподобия

- Тогда вероятности наблюдаемых частот:
 - $- L(H_1) = b(c_{12}; c_1, p)b(c_2 - c_{12}; N - c_1, p)$
 - $- L(H_2) = b(c_{12}; c_1, p_1)b(c_2 - c_{12}; N - c_1, p_2)$

	H_1 (независимость)	H_2 (зависимость)
$P(w_2 w_1)$	$p = \frac{c_2}{N}$	$p_1 = \frac{c_{12}}{c_1}$
$P(w_2 \neg w_1)$	$p = \frac{c_2}{N}$	$p_2 = \frac{c_2 - c_{12}}{N - c_1}$
c_{12} из c_1 биграмм – это w_1w_2	$b(c_{12}; c_1, p)$	$b(c_{12}; c_1, p_1)$
$c_2 - c_{12}$ из $N - c_1$ биграмм – это $\neg w_1w_2$	$b(c_2 - c_{12}; N - c_1, p)$	$b(c_2 - c_{12}; N - c_1, p_2)$

Отношение правдоподобия

- Выражаем отношение правдоподобия:

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

- Если $-2 \log \lambda >$ порога – отвергаем гипотезу о независимости

Отношение правдоподобия

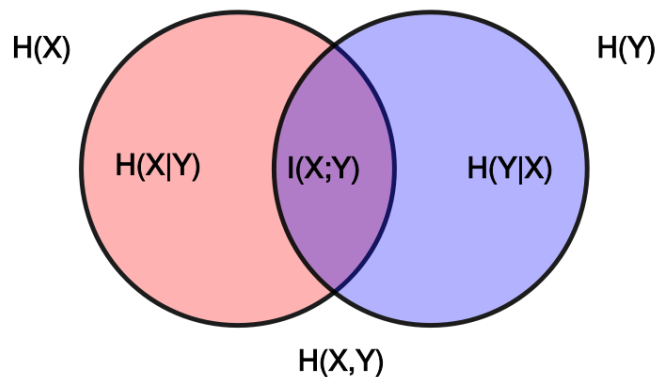
	$-2\log \lambda$	C1	C2	C12
мощный россия	66.800607	14709	513837	2
мощный землетрясение	58.521963	14709	5490	543
мощный взрыв	44.635622	14709	20928	930
мощный работа	19.562579	14709	242103	4
мощный город	15.886387	14709	212631	4
мощный решение	15.403840	14709	185295	3
мощный матч	14.409602	14709	222457	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_1

С достоверностью 0.005, можно принять H_1

Взаимная информация

- Говорит, сколько информации несет в себе одно слово относительно другого



Для биграмм интересно отношение:

$$PMI(X;Y) = \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

$H(X) = - \sum_{i=1}^n p_i \log_2 p_i$ - энтропия

$H(X|Y) = H(X;Y) - H(X)$ - условная энтропия

$H(X;Y) = - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y)$ - взаимная

Взаимная информация:

$$I(X;Y) = H(X) - H(X|Y) =$$

$$\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Топ биграмм (pmi)

	pmi	m1	C1	C2	C12
мощный 23мр	12.629566	1.354962e-07	14709	4	2
мощный медиаимперия	11.822211	1.268345e-07	14709	7	2
мощный дожимной	11.542103	2.476587e-07	14709	17	4
мощный иммунодепрессант	11.542103	2.476587e-07	14709	17	4
мощный нейротоксин	11.492063	3.082312e-07	14709	22	5
мощный антипиар	11.459641	2.458893e-07	14709	18	4
мощный плавник	11.451229	1.167117e-06	14709	86	19
мощный жиросжигатель	11.381639	2.442156e-07	14709	19	4
мощный излияние	11.214529	1.804724e-07	14709	16	3
мощный электроразряд	11.170135	1.198387e-07	14709	11	2
мощный 21мр	11.044604	1.777379e-07	14709	18	3
мощный кусачки	10.951494	2.937325e-07	14709	32	5
мощный пиропатрон	10.892601	1.752917e-07	14709	20	3
мощный айпад	10.307638	1.658781e-07	14709	30	3
мощный видеоадаптер	10.307638	1.105854e-07	14709	20	2

Оценка PMI. Разреженность

$$I(x, y) = \log \frac{P(xy)}{P(x)P(y)} = \log \frac{P(x)}{P(x)P(y)} = \log \frac{1}{P(y)} \quad \text{Полная зависимость}$$

$$I(x, y) = \log \frac{P(xy)}{P(x)P(y)} = \log \frac{P(x)P(y)}{P(x)P(y)} = \log 1 = 0 \quad \text{Полная независимость}$$

Оценка PMI. Выводы

- PMI плохо подходит для извлечения коллокаций, т.к. плохо работает в случае сильно разреженных данных
- PMI хорошо подходит для проверки гипотезы о независимости слов

N-ГРАММЫ

N-граммы

- N-грамма (n-gram) – последовательность из N слов:
 - N=1 – униграмма (unigram)
 - N=2 – биграмма (bigram)
 - N=3 – триграмма (trigram)

N-граммная модель

- Обучается на каком-то корпусе
- Предсказывает вероятности N-грамм:

$$p(w_1, w_2, \dots, w_n)$$

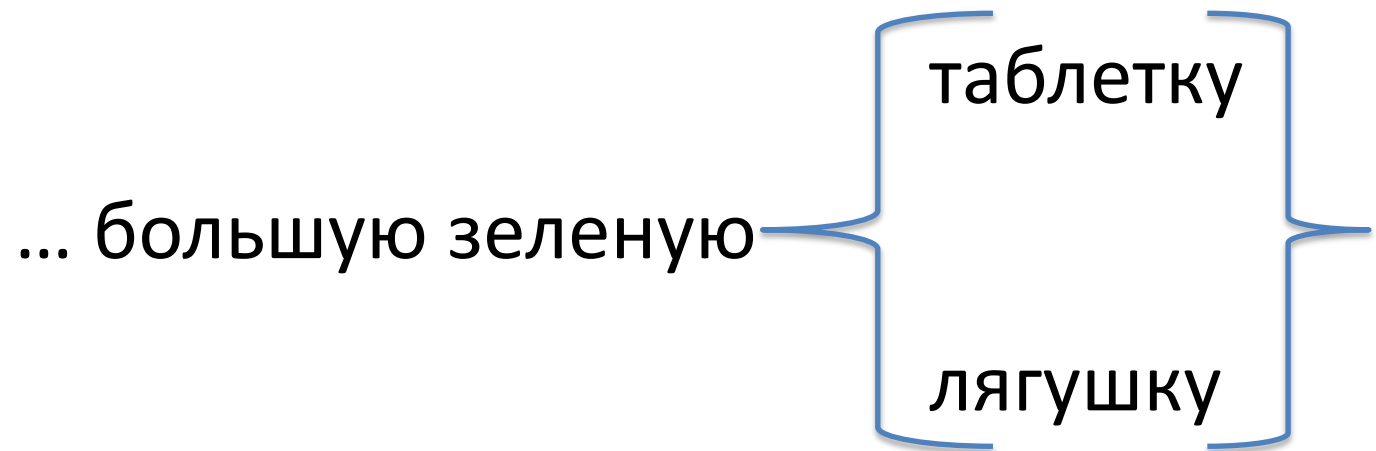
- Предсказывает вероятность появления следующего слова в зависимости от предыдущих:

$$p(w_n | w_1, w_2, \dots, w_{n-1})$$

N-граммная модель


- Порядок предшествующих слов не важен!
- Когда мы считаем частоты и вероятности, то считаем все последовательности, отличающиеся только порядком слов, одной и той же N-граммой

Как это работает:



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.
4-х граммная	$20\,000^3 \times 19\,999 = 1,6 \times 10^{17}$

Как быть:

- Уменьшать N
- Делать классы эквивалентности (леммы, синонимы, ...)

Обучение N-граммных моделей

- Как нам оценить параметры модели?
- У нас задача машинного обучения:
 - Оцениваем параметры на обучающей выборке
 - Оцениваем качество модели на отложенной тестовой выборке

Обучение N-граммных моделей

- Используем оценку максимального правдоподобия (MLE)
- Оценка вероятности N-граммы:

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

- Оценка вероятности N-го слова:

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

Пример: 3-граммная модель

(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

Пример: оценка фразы

- Оценим фразу

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

моделями:

- Unigram
- Bigram
- Trigram

Пример: 1-граммная модель

	президент	сша	барак	обама	решить	сняться	в	телепередача	с	бears	грилсом
count	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
- В топ лезут стоп-слова

Топовые слова по вероятностям
в - 0.0433
и - 0.0245
на - 0.0194

Пример: 2-граммная модель

	президент сша	сша барак	барак обاما	обاما решить	решить сняться	сняться в	в телепередача	телепередача с	с беар	беар грилсом
count	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 вероятности биграмм со словом президент

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

В итоге

- Unigram – полностью игнорирует контекст, но это может быть полезно для общих слов
- Bigram – использует предыдущее слово для оценки вероятности, получаем лучшую модель
- 3-gram модель должна работать хорошо, но... На практике появляется очень много дырок в оценке вероятности из-за разреженности данных!

Как быть с разреженными данными

- Подбираем правильный размер модели
 - Мало данных – снижаем n
 - Много данных – повышаем n
- Сглаживание!

Сглаживание Лапласа (adding one)

- Самый простой вид сглаживания

$$p_{lap}(w_1 \dots w_n) = \frac{C(w_1 \dots w_n) + 1}{C(w_1 \dots w_{n-1}) + B}$$

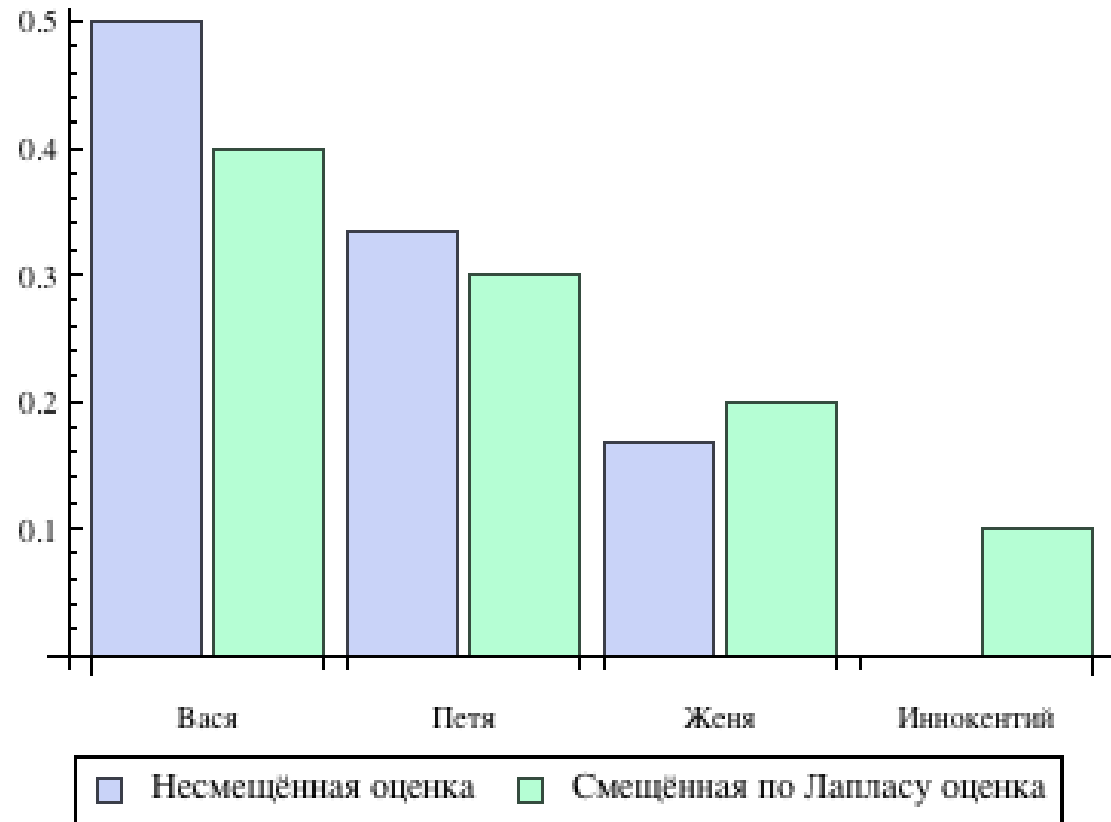
тут B – размер словаря

- Есть недостатки:
 - Провоцирует сильную погрешность
 - Иногда несглаженная модель показывает лучшие результаты

Сглаживание Лапласа: пример

Имя	Частота
Вася	3
Петя	2
Женя	1

Неизвестное слово:
Иннокентий



Применение в поиске

- Оценка части запроса на вхождение известных пассажей

[где приобрести бесплатно] [георгиевская ленточка]

p(где <s>)	= [2gram] 0.0276706
p(приобрести где ...)	= [3gram] 0.000907595
p(бесплатно приобрести ...)	= [3gram] 0.00042164
p(георгиевская бесплатно ...)	= [1gram] 6.14581e-07
p(ленточка георгиевская ...)	= [2gram] 0.568604
p(</s> ленточка ...)	= [3gram] 0.20317

МАРКОВСКИЕ МОДЕЛИ

Цепи Маркова



Doudou sleeping



Doudou eating



Doudou training



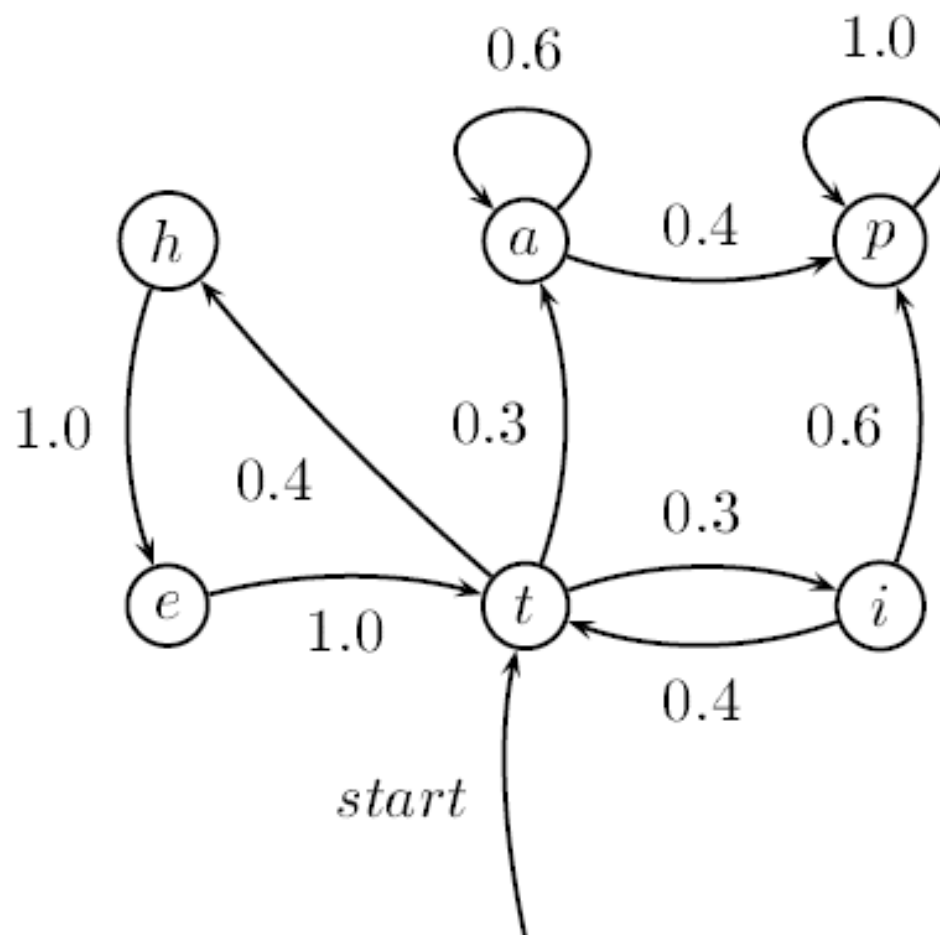
Цепи Маркова

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



Последовательность состояний

Вероятность последовательности:

- $$P(X_1, \dots, X_T) =$$
$$P(X_1)P(X_2|X_1)P(X_T|X_1, \dots, X_{T-1}) =$$
$$P(X_1)P(X_2|X_1)P(X_T|X_{T-1}) =$$
$$\pi_{x_1} \prod_{t=1}^T a_{X_t X_{t+1}}$$

Вероятность последовательности:

- $$P(X_1, \dots, X_T) =$$
$$P(X_1)P(X_2|X_1)P(X_T|X_1, \dots, X_{T-1}) =$$
$$P(X_1)P(X_2|X_1)P(X_T|X_{T-1}) =$$
$$\pi_{x_1} \prod_{t=1}^T a_{X_t X_{t+1}}$$

- $$P(t, i, p) = P(t)P(X_2 = i | X_1 =$$
$$t)P(X_3 = p | X_2 = t) = 1.0 \times 0.3 \times 0.6 = 0.18$$

- $$P(t, i, p) = P(t)P(X_2 = i | X_1 =$$
$$t)P(X_3 = p | X_2 = t) = 1.0 \times 0.3 \times 0.6 = 0.18$$

Подсчитаем вероятность

“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

“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	$p_s(\text{сняться} \rightarrow v)$
0.000018	$p_s(v \rightarrow \text{телепередача})$
0.014085	$p_s(\text{телепередача} \rightarrow c)$
0.000062	$p_s(c \rightarrow \text{беар})$
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*

Применение Марковских цепей

- Оцениваем авторство документа. Каждый автор имеет свой стиль
- Случайное блуждание по Интернету (Page Rank)
- Ранжирование с помощью языковых моделей в поиске

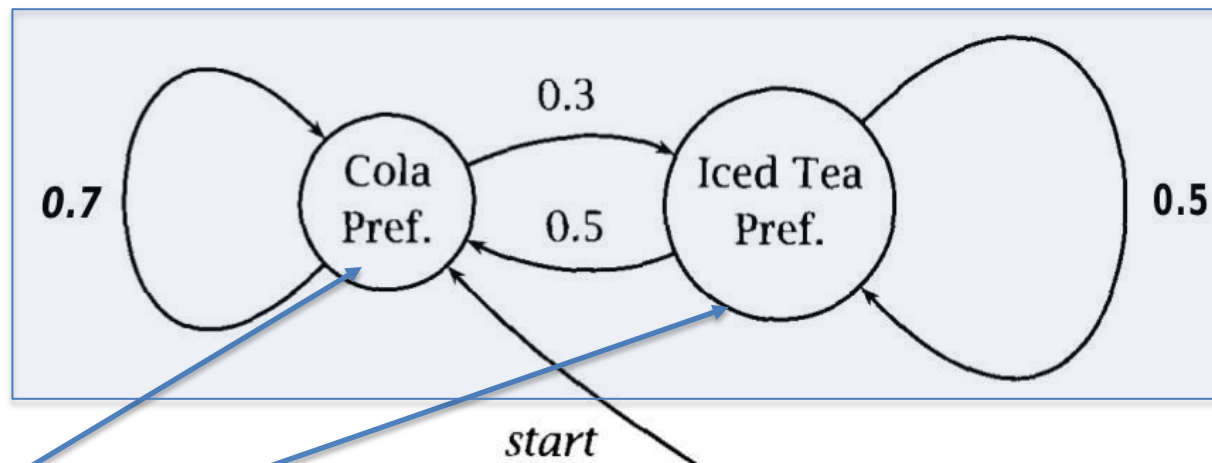
Применение в поиске

- Языковые модели можно использовать в ранжировании:
 - Строим языковую модель для каждого из документов в коллекции
 - С помощью этих моделей предсказываем вероятности текущего запроса
 - Ранжируем документы в порядке убывания таких вероятностей
 - Сглаживаем разреженные данные

Автомат прохладительных напитков



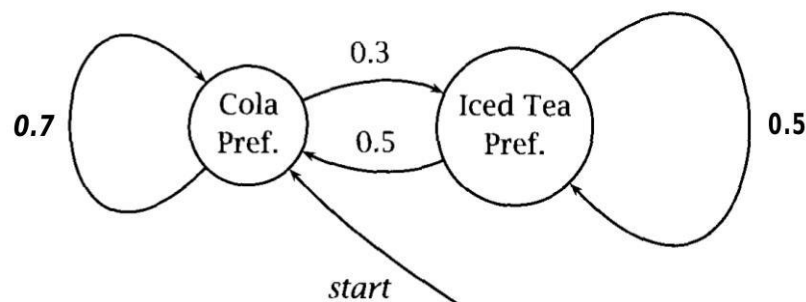
WWW.3DVSEM.COM



Два состояния	Кратко
Cola Pref.	CP
Iced Tea Pref.	IP

Состояния	Разливаемые напитки		
	Cola	Iced Tea (ice_t)	Lemonade (lem)
CP	0.6	0.1	0.3
IP	0.1	0.7	0.2

Пример



Состояния	Разливаемые напитки		
	Cola	Iced 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 автомат стартует в состоянии CP?

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

Скрытая Марковская модель

- Модель проходит через некую скрытую последовательность внутренних состояний, о которых можно судить только по наблюдаемым параметрам («символам»)
- При переходе из состояния s_i в состояние s_j модель генерит “символ” с вероятностью:

$$P(O_t = k | X_t = s_i, X_{t+1} = s_j) = b_{ijk}$$

Основные термины

- $S = \{s_1, \dots, s_N\} = \{1 \dots N\}$ – состояния модели
- $K = \{k_1, \dots, M\}$ – набор возможных символов
- $\Pi = \{\pi_i\}, i \in S$ – вероятности начальных состояний
- $A = \{a_{ij}\}, i, j \in S$ – вероятности перехода
- $B = \{b_{ijk}\}, i, j \in S, k \in K$ – вероятность “символа”
- $X = (X_1, \dots, X_{T+1})$ – последовательность состояний
- $O = (o_1, \dots, o_T)$ – выходная последовательность

Три задачи НММ

1. Дано: модель $\mu(A, B, \Pi)$ и наблюдения O .
 - Оценить насколько вероятны наши наблюдения, т.е. оценить вероятность $P(O|\mu)$
2. Дано: модель μ и наблюдения O .
 - Выбрать последовательность (X_1, \dots, X_{T+1}) , которая лучше всего описывает наши наблюдения
3. Дано: последовательность наблюдений O и набор моделей с вариацией параметров $\mu(A, B, \Pi)$
 - Найти модель, лучше всего описывающую наши наблюдения

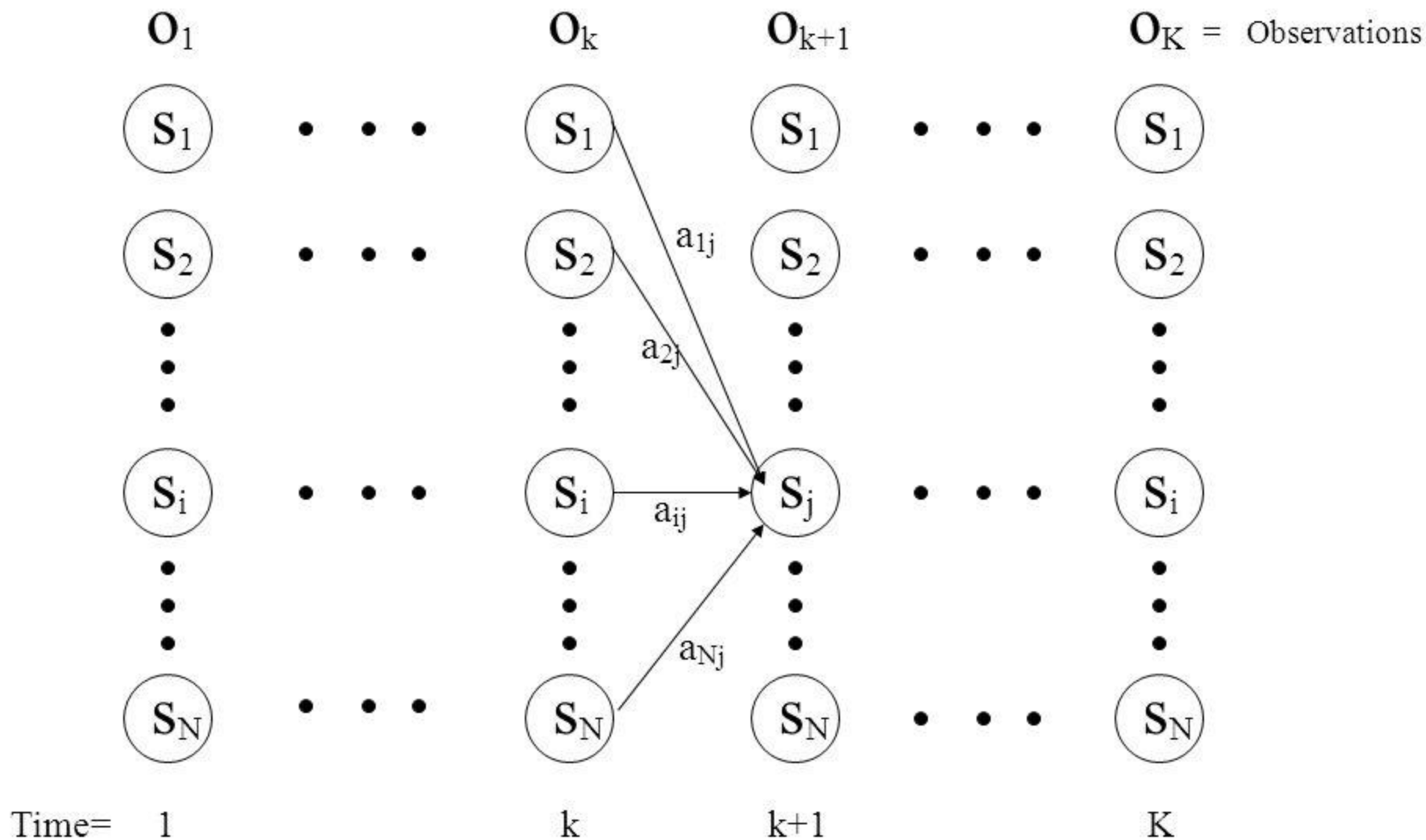
Задача 1

- Мы могли бы решить первые две задачи, если бы знали совместную вероятность $P(O, X|\mu)$
- Из нее легко получить полные вероятности $P(O|\mu)$ и $P(X|\mu)$
- А еще условные вероятности $P(O|X, \mu)$ и $P(X|O, \mu)$

Задача 1

- $P(O|\mu) = \sum_X P(O, X|\mu) P(X|\mu)$
- $P(O, X|\mu) = P(O|X, \mu)P(X|\mu)$
- $P(O|X, \mu) = \prod_{t=1}^T P(o_t|X_t, X_{t+1}, \mu) =$
 $= b_{x_1 x_2 o_1} b_{x_1 x_2 o_2} \dots b_{x_T x_{T+1} o_T}$
- $P(X|\mu) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3} \dots a_{x_T x_{T+1}}$
- $P(O|\mu) = \sum_{x_1 \dots x_{T+1}} \pi_{x_1} \prod_{t=1}^T a_{x_t x_{t+1}} b_{x_t x_{t+1} o_t}$
- Получаем $O((T + 1) N^{T+1})$ перемножений
- Выход – динамическое программирование

Trellis representation of an HMM



Задача 1. Прямой проход

- Примем, что:

$$\alpha_i(t) = P(o_1 o_2 \dots o_{t-1}, X_t = i | \mu)$$

- Инициализация:

$$\alpha_i(1) = \pi_i, 1 \leq i \leq N$$

- Индукция:

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

- Итого:

$$P(O | \mu) = \sum_{i=1}^N \alpha_i(T+1)$$

Задача 1. Обратный проход

- Примем, что:

$$\beta_i(t) = P(o_t \dots o_T, X_t = i | \mu)$$

- Инициализация:

$$\beta_i(T + 1) = 1, 1 \leq i \leq N$$

- Индукция:

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

- Итого:

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

Задача 1. Итоги

- С помощью прямого или обратного проходов мы можем решить задачу с использованием $2N^2T$ умножений

Задача 2

- Находим последовательность состояний, которые лучше описывает наблюдения
- Т.н. «задача декодирования»
- Что значит лучше?
- Будем искать путь X , который максимизирует $P(X|O, \mu)$

Алгоритм Витерби

- Ищем $\arg \max_X P(X|O, \mu)$
- Достаточно максимизировать $\arg \max_X P(X, O|\mu)$ при фиксированном O
- $\delta_j(t) = \max_{x_1 \dots x_{t-1}} P(x_1 \dots x_{t-1}, o_1 \dots o_{t-1}, X_t = j|\mu)$
- Инициализируем:
 - $\delta_j(t) = \pi_j, 1 \leq j \leq N$
- Индукция:
 - $\delta_j(t+1) = \max_{1 \leq i \leq N} \delta_j(t) a_{ij} b_{ijo_t}, 1 \leq j \leq N$
 - Сохраняем состояние: $\psi_j(t+1) = \operatorname{argmax} \delta_j(t) a_{ij} b_{ijo_t}, 1 \leq j \leq N$
- Заканчиваем, и считываем, что получилось:
 - $\hat{X}_{T+1} = \operatorname{argmax} \delta_i(T+1), 1 \leq i \leq N$ - финал
 - $\hat{X}_t = \psi_{\hat{X}_{t+1}}(t+1)$
 - $P(\hat{X}) = \max \delta_i(T+1), 1 \leq i \leq N$
- Возвращаем список лучших состояний

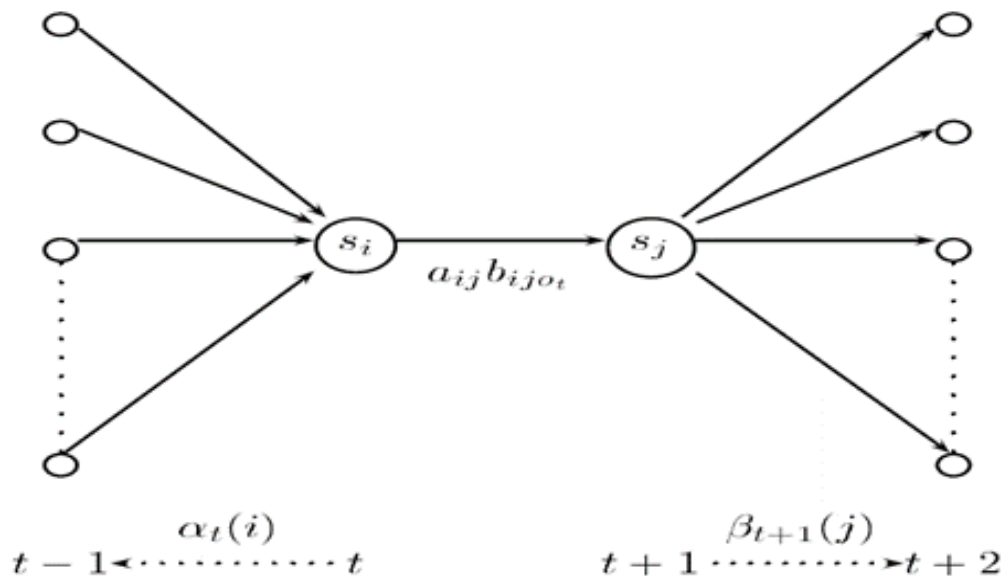
Задача 3

- Ищем оптимальные значения модельных параметров: $\mu(A, B, \pi)$
- Используем оценку максимального правдоподобия для максимизации:
$$P(O|\mu) = \arg \max_{\mu} P(O_{train} | \mu)$$
- Нет аналитического решения!
- Используем итеративный алгоритм Баума-Велша (Baum-Welch)

Алгоритм Баума-Велша

- Начинаем с некоторой модели μ (напр. случайной)
- Прогоняем O через модель и оцениваем ожидание каждого из параметров модели
- Меняем модель так, что максимизировать вклад часто используемых путей
- Повторяем до сходимости

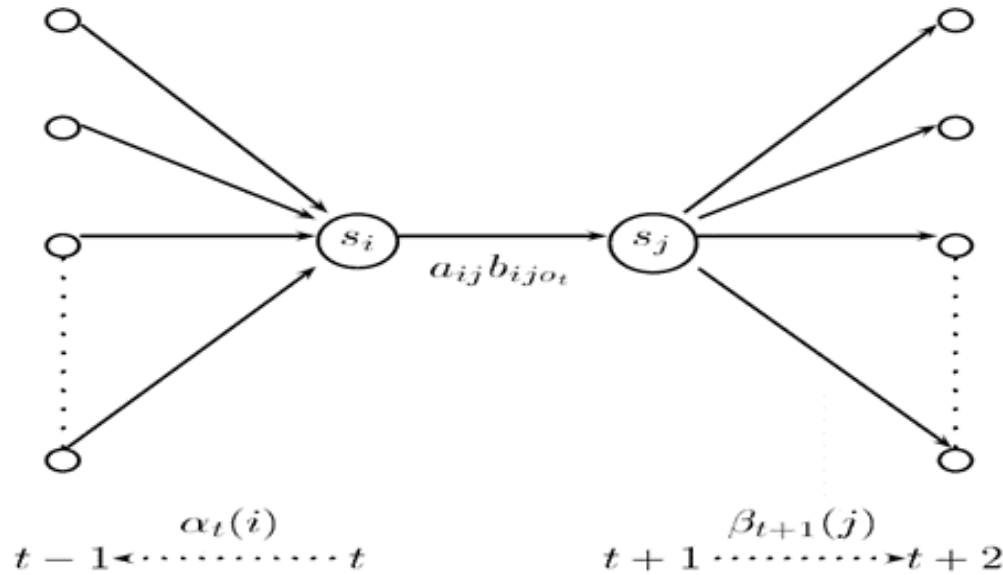
Вероятности



- Вероятность нахождения модели в состоянии i для наблюдаемой последовательности O :

$$\gamma_t(i) = P(X_t = i | O, \mu) = \frac{P(X_t = i, O | \mu)}{P(O | \mu)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^N \alpha_j(t)\beta_j(t)}$$

Вероятности



- Определим совместную вероятность нахождения в состоянии i и j в момент времени t и $t+1$

$$\begin{aligned}
 p_t(i, j) &= P(X_t = i, X_{t+1} = j \mid O, \mu) = \frac{P(X_t = i, X_{t+1} = j, O \mid \mu)}{P(O \mid \mu)} \\
 &= \frac{\alpha_i(t) a_{ij} b_{ij o_t} \beta_j(t+1)}{\sum_{m=1}^N \alpha_m(t) \beta_m(t)} = \frac{\alpha_i(t) a_{ij} b_{ij o_t} \beta_j(t+1)}{\sum_{m=1}^N \sum_{n=1}^N \alpha_m(t) a_{mn} b_{mn o_t} \beta_n(t+1)}
 \end{aligned}$$

Оценка параметров

- Теперь суммируем по временному индексу:

$\sum_{t=1}^T \gamma_i(t)$ -- оценка количества переходов из состояния i для всех наблюдений O

$\sum_{t=1}^T p_t(i, j)$ -- оценка количества переходов из состояния i в j для всех наблюдений O

Оценка параметров

- Пере-оцениваем параметры модели:

$\hat{\pi}_i = \gamma_i(1)$ – ожидаемая частота в состоянии i в момент $t=1$

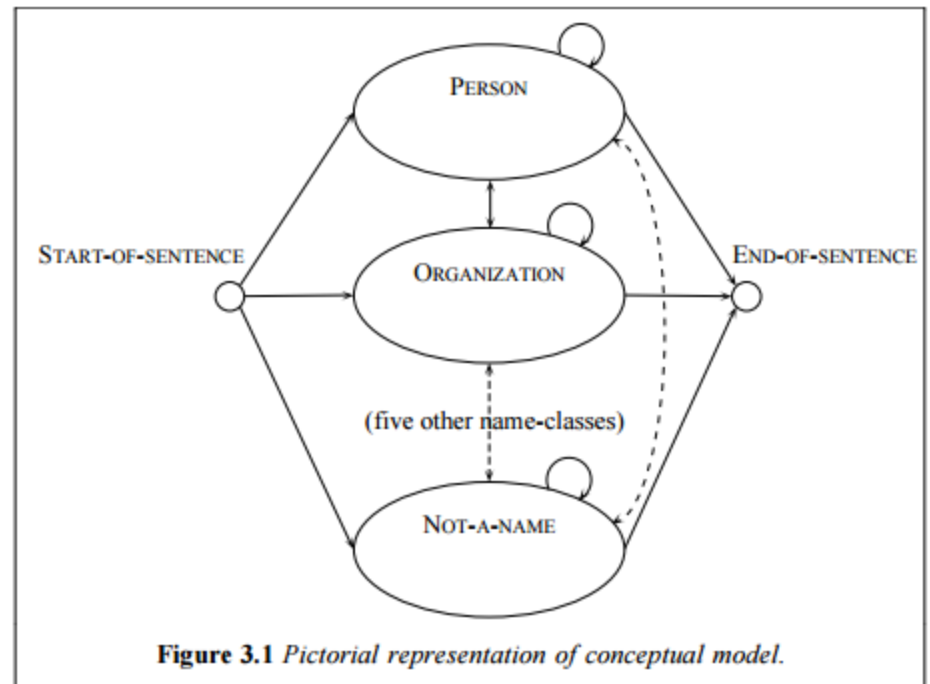
$$\hat{a}_{ij} = \frac{\text{ожидаемое кол-во переходов из } i \text{ в } j}{\text{ожидаемое кол-во переходов из } i}$$

$$\hat{b}_{ijk} = \frac{\text{ожидаемое кол-во переходов из } i \text{ в } j \text{ при } k \text{ набл.}}{\text{ожидаемое кол-во переходов из } i \text{ в } j}$$

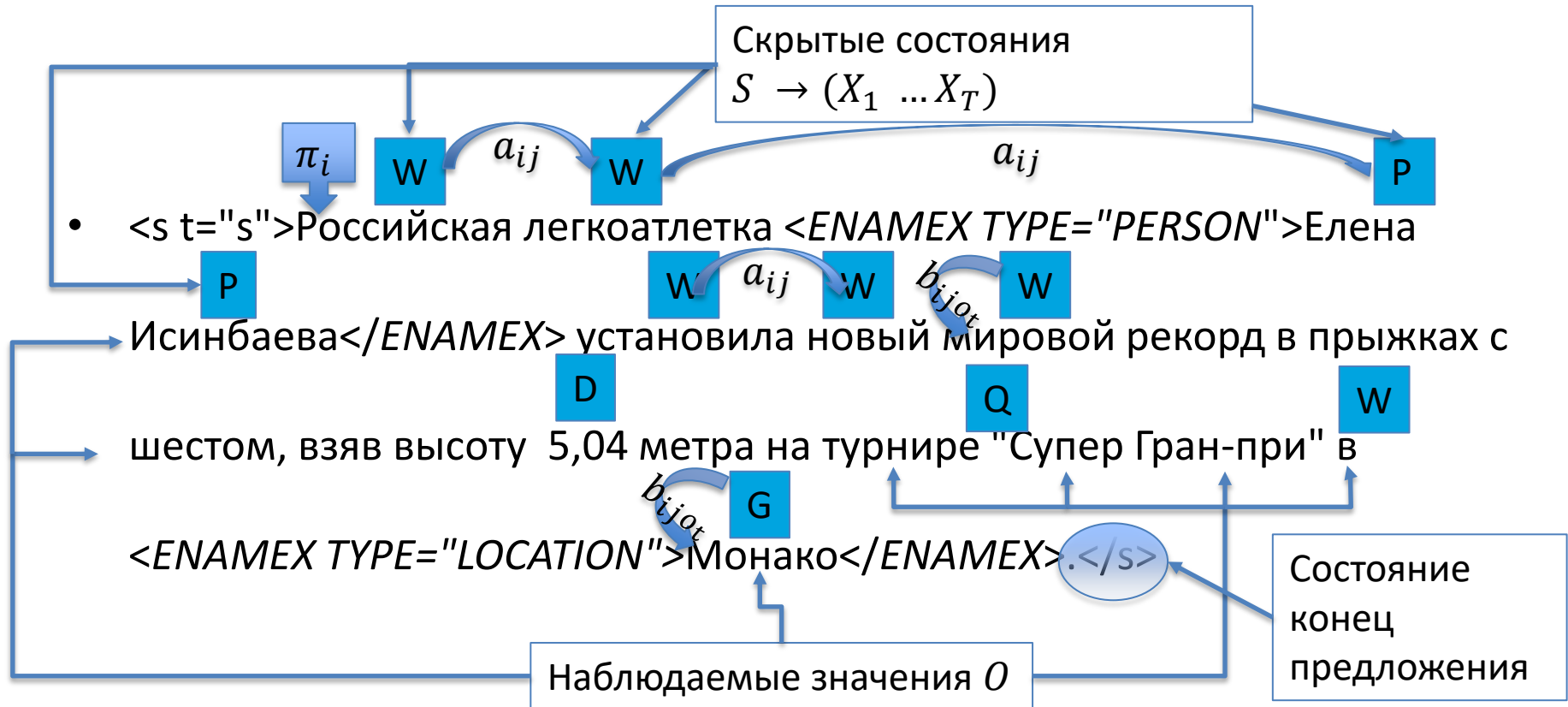
Тегирование

Концептуальная схема модели

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	,



Пример: NER



Используем алгоритм Витерби для декодирования скрытой последовательности по нашим наблюдениям

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

Вопросы?