N-gram Language Models

Занятие 3

Маленький эксперимент

Он шел и курил …

Маленький эксперимент

Он шел и курил сигарету

Он шел и курил трубку

Он шел и курил.

Что вероятнее?

Из рубрики Капитан очевидность

Он шел и курил сигарету

Шел он и сигарету курил

Он сигарету и шел курил

Очевидно людям, совсем не очевидно машине.

Вероятность

Интуитивно кажется, что одно предложение **вероятнее** других.

Задача: приписывать **вероятность** фразам/предложениям.

Из этого вытекает другая задача:

предсказывать следующее слово.

Зачем это нужно?

- Speech recognition
- Генерация речи
- Машинный перевод
- Саммаризация
- QA
- Спелл-чекинг:
    - я гтовлю лекцию VS я готовлю лекцию
    - подай мне нос VS подай мне нож

Языковая модель

aka LM

Языковая модель — модель, которая приписывает вероятность фрагменту текста (высказыванию, предложению…)

(Определение грубое, зато понятное)

Языковая модель
aka LM

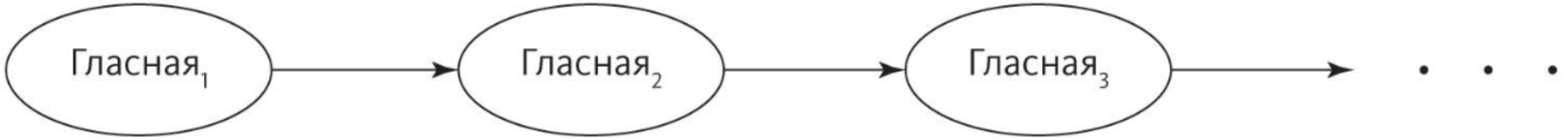
Иными словами:

- максимизирует вероятность реальных текстов
- минимизирует вероятность нереальных текстов

Цепь Маркова

Работа А. Маркова «Пример статистического исследования над текстом “Евгения Онегина” иллюстрирующий связь испытаний в цепь», 1913 год.

Допущение Маркова: вероятность появления той или иной буквы зависит от буквы, непосредственно ей предшествующей.



Цепь Маркова

Предположение: вероятность наступления события зависит только от предыдущего состояния.

Общая философия: мы можем предположить будущее по настоящему, сильно не углубляясь в прошлое.

Цепь Маркова

Допущение Маркова: вероятность появления той или иной буквы зависит от буквы, непосредственно ей предшествующей.

Применительно к тексту:

Следующее слово зависит только от предыдущего (N предыдущих)

Вероятность языковых событий

- Вероятность основана на подсчете событий (частотность)
- Мы считаем вероятность языковых явлений по корпусу

Немного теории

Классическая вероятность

Вероятностью наступления события A в некотором испытании называют отношение

$$P(A) = \frac{m}{n}$$

n – общее число всех равновозможных, элементарных исходов этого испытания, которые образуют полную группу событий;
m – количество элементарных исходов, ***благоприятствующих*** событию A.

Попробуем сами

Вероятность появления слова *однокомнатная* в НКРЯ

однокомнатная - 69

всего слов в корпусе НКРЯ - 337 025 184 слова.

$$P(\text{однокомнатная}) = \frac{freq(\text{однокомнатная})}{freq(\text{НКРЯ})} = \frac{69}{337025184} = 0.000000204732474828943 35$$

N-граммы

N последовательно стоящих друг за другом слов.

униграммы — — Однокомнатная квартира в Москве.

биграммы — — <s> Однокомнатная квартира в Москве. </s>

триграммы — — <s> <s> Однокомнатная квартира в Москве. </s> </s>

N-граммы

N последовательно стоящих друг за другом слов.

униграммы — — Однокомнатная квартира в Москве

биграммы — — <s> Однокомнатная квартира в Москве </s>

триграммы — — <s> <s> Однокомнатная квартира в Москве </s> </s>

Скользящее окно

Униграммная модель

При $n = 1$ (униграммная модель) вероятности слов соответствуют частотам слов в корпусе.

НКРЯ - 337 025 184 слова.

однокомнатная - 156 вхождений

квартира - 9304 вхождения

Униграммная модель

При $n = 1$ (униграммная модель) вероятности слов соответствуют частотам слов в корпусе.

НКРЯ - 337 025 184 слова.

однокомнатная - 156 вхождений

квартира - 9304 вхождения

$$\frac{\text{однокомнатная}}{\text{НКРЯ}} * \frac{\text{квартира}}{\text{НКРЯ}} = \frac{156}{337025184} * \frac{9304}{337025184} = 0.000000000012778197347598422$$

Униграммная модель

Мы посчитали вероятность встретить фразу «однокомнатная квартира» в корпусе НКРЯ

Проблемы такой модели?

Униграммная модель

Униграммная модель не учитывает порядок слов и их сочетаемость.

Но ведь в языке слова не случайно встречаются друг с другом, верно?

Биграммная модель

Учитывает вероятности переходов из предыдущего слова в текущее слово

То есть уже учитывает **контекст**

Биграммная модель

однокомнатная квартира

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(\text{квартира}|\text{однокомнатная}) = \frac{freq(\text{однокомнатная квартира})}{freq(\text{однокомнатная})} = \frac{69}{156} = 0.44$$

Вероятность гораздо красивее. И в ней больше смысла!

Триграммная модель

Вероятность слова, учитывая два предыдущих

Посчитайте вероятность, что после фразы *однокомнатная квартира* будет идти предлог «*в*» в триграммной модели по корпусу НКРЯ

Триграммная модель

Вероятность слова, учитывая два предыдущих

$$P(\text{в}|\text{однокомнатная квартира}) = \frac{freq(\text{однокомнатная квартира в})}{freq(\text{однокомнатная квартира})} = \frac{14}{69} = 0.2$$

To sum up

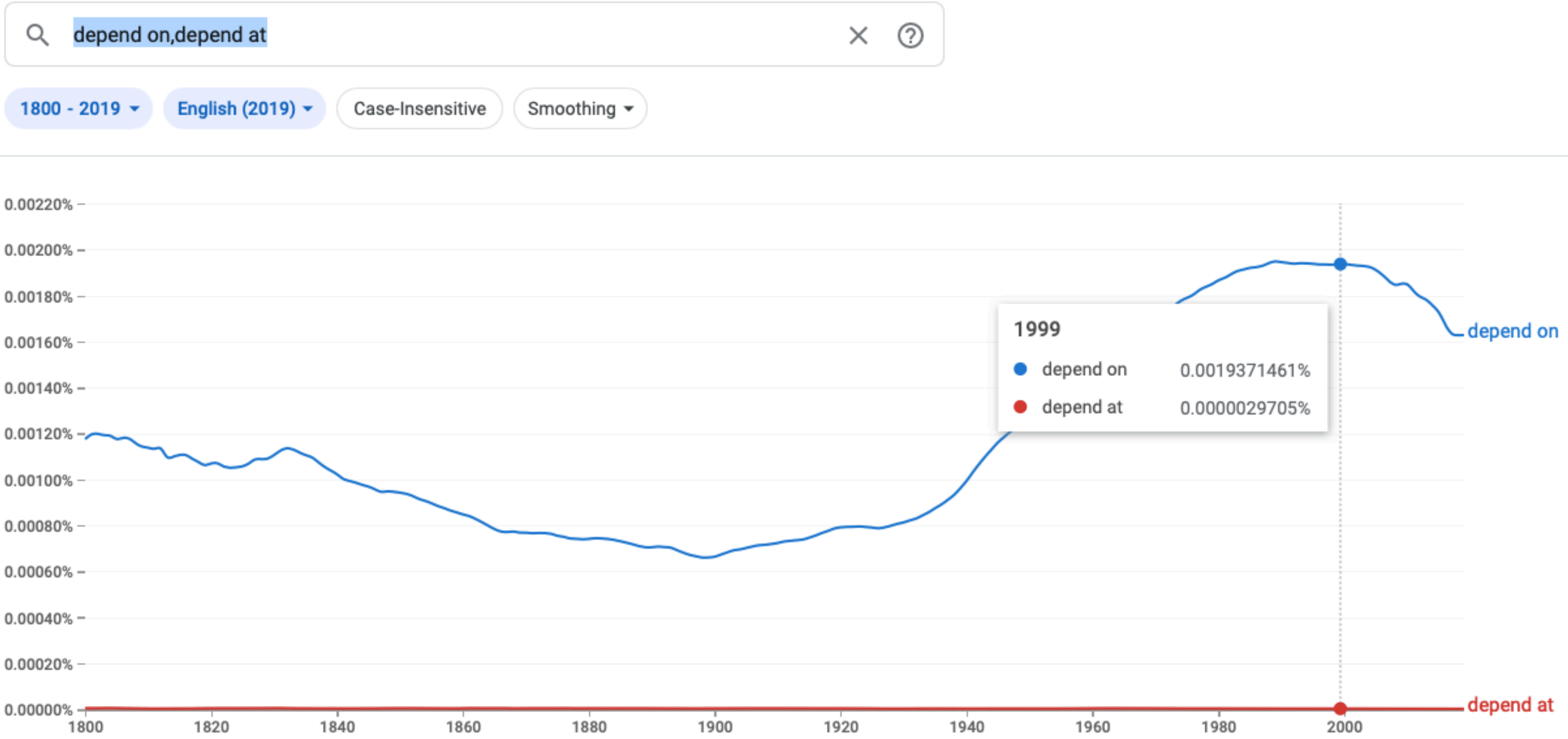
Предположение N-граммной модели: вероятность слова зависит от n - 1 предыдущего.

$$p(w_N | w_1 \dots w_{N-1}) = p(w_N | w_{N-n+1} \dots w_{N-1})$$

Формула

$$P(w_{1:n}) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

Google N-gram Viewer



Вероятность предложения

С помощью N-граммной модели можно считать не только вероятность фраз, но и вероятности целых предложений.

Посчитайте вероятность предложения в биграммной модели

Где ты живешь

Подсказка: нужно учитывать начало и конец предложения

Вероятность предложения

Биграммная модель

<s> где - 41 268 вхождений

<s> - 28073538 (число предложений в корпусе)

где ты - 4378 вхождений

где - 370823 вхождения

ты живешь - 989

ты - 613676

живешь </s> - 0 !!!

живешь - 4365

Вероятность предложения

Биграммная модель

Вся вероятность превратится в 0, потому что *живешь* ни разу не встретилось как последнее слово в предложении.

Нулевые вероятности

я читал		1864			
я читал	книгу	19	$\frac{19}{1864}$	$\approx$	0.010
я читал	газету	3	$\frac{3}{1864}$	$\approx$	0.002
я читал	лекцию	11	$\frac{11}{1864}$	$\approx$	0.006
я читал	доклад	0	$\frac{0}{1864}$	$=$	0?
я читал	инструкцию	0	$\frac{0}{1864}$	$=$	0?

Нулевые вероятности

Out-of-vocabulary (OOV)

- сленг
- профессионализмы
- проф. сленг
- неологизмы
- окказионализмы
- намеренное искажение

Нулевые вероятности

Out-of-vocabulary (OOV)

Объём всего корпуса: 126 901 документ, 337 025 184 слова.

"**захардкодить**"

По этому запросу ничего не найдено.

"**суперлатив**"

По этому запросу ничего не найдено.

"**править**"
на расстоянии **1** от "**баги**"

По этому запросу ничего не найдено.

"**кринж**"

По этому запросу ничего не найдено.

Омонимия

Out-of-vocabulary (OOV)

"тяночка"

Найдено: 1 документ, 2 вхождения.

Распределение по годам Статистика 1-граммы 2-граммы 3-граммы 4-граммы 5-граммы

Поискать в других корпусах: газетном, региональном, диалектном, поэтическом, устном, акцентологическом, мультимедийном, мультип

Страницы: **1**

1. **В. Я. Зазубрин. Горы (1934)** [омонимия не снята] **Все примеры (2)**

   — Проходу он мне, **Тяночка**, не дает, лапает. [В. Я. Зазубрин. Горы (1934)] [омонимия не снята] ←...→

   Ему отец сколь разов говорил: «**Тяночка**, не в ту сторону тянешь». [В. Я. Зазубрин. Горы (1934)] [омонимия не снята] ←...→

Сглаживание

Как избавиться от нулей

- Сглаживание Лапласа aka Add-one smoothing
- Add-k smoothing
- Откат (backoff)
- Интерполяция

Сглаживание Лапласа

add-one smoothing

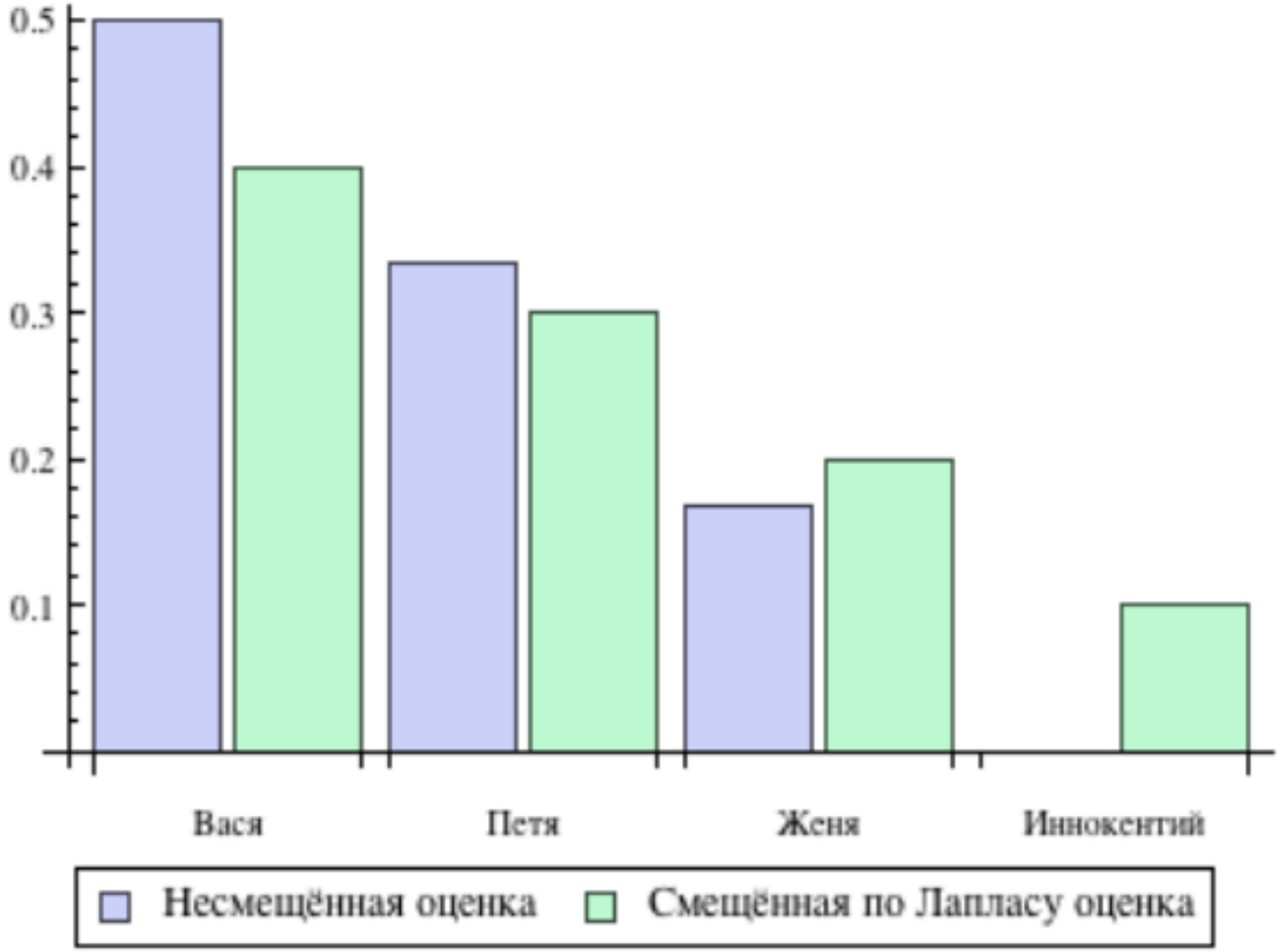
К вероятности каждой N-граммы прибавлять 1, чтобы избавиться от нулей

И в делитель придется добавить V - кол-во типов токенов (уникальных токенов) в словаре.

$$P^*_{\text{Laplace}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

Сглаживание Лапласа

Имя	Частота
Вася	3
Петя	2
Женя	1



Сглаживание Лапласа

+ Нули убрал

НО

- Сильно меняет вероятность

- Скашивается в сторону редких слов за счет вероятности более частотных

Add-k smoothing

Прибавлять не единицу, а что-то поменьше (0.5, 0.1, 0.05, …)

$$P^*_{\text{Add-k}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n) + k}{C(w_{n-1}) + kV}$$

- Лучше, чем сглаживание Лапласа
- Доказали, что все равно неприятно искажает вероятности.

Откат

Backoff

- Нулевая вероятность у триграммы? Посчитать биграмму!
- Нулевая вероятность у биграммы? Посчитать униграмму!

Т.е. откатываемся по N (N - 1, N - 2, …, N - k)

Интерполяция

Не заменяем триграмму на биграмму, если у триграммы P = 0,

а всегда суммируем вероятности N-грамм

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n)$$

при этом $\sum_i \lambda_i = 1$

Сглаживание

to sum up

- От нулей избавиться можно, но сложно не исказить вероятность
- Есть методы проще, есть методы сложнее
- Мы посмотрели далеко не все

N-граммная языковая модель

Выводы

- Самая простая языковая модель
- Сложные нейросетевые модели BERT (SOTA) и ELMo, которые используют в индустрии NLP - тоже языковые модели! (Т.е. тоже приписывают вероятности)
- Рассматриваем н-граммы, потому что это база
- И это можно посчитать ручками!