Большие языковые модели, трансформеры, ChatGPT

Лукашевич Н.В., в.н.с. НИВЦ МГУ, профессор ВМК МГУ

Тихомиров М.М., н.с. НИВЦ МГУ

Чернышев Д.И., аспирант ВМК

Диалог с ChatGPT в качестве эпиграфа

ZO

сегодня я случайно села на свою любимую свигму и у нее отвалился хрюксер. что делать?

Если у вас отвалился хрюксер со свигмы, то следует немедленно обратиться за помощью к квалифицированному специалисту. Хрюксер является важным элементом свигмы, который обеспечивает ее стабильность и безопасность. Его отсутствие может привести к тому, что свигма станет нестабильной и небезопасной для использования.

Свяжитесь с магазином или производителем свигмы, чтобы узнать, как произвести ремонт хрюксера или приобрести новый. В случае необходимости может потребоваться профессиональное установление хрюксера, так что убедитесь, что у вас есть доступ к квалифицированному специалисту, который сможет выполнить эту работу. Пока хрюксер не будет заменен или установлен, не следует использовать свигм

План по дням

- 1. Нграммные языковые модели (Лукашевич Н.В., Тихомиров М.М.)
  - Основы нграммных языковых моделей
  - Исправление опечаток
  - Метод word2vec для векторизации слов

- 2.Автоматическое аннотирование (суммаризация текстов) (Чернышев Д.И.)
  - Векторные представления документов
  - Методы автоматического аннотирования
  - Архитектура типа трансформер

- 3. Информационный поиск (Тихомиров М.М)
  - Классические модели
  - Модели на основе нейронных сетей
  - Методы тестирования
- 4. Модели семейства GPT (Тихомиров М.М., Лукашевич Н.В.)

N-граммные модели

Языковые статистические модели

Языковые модели (language models)

- Определение вероятности предложений, последовательностей слов

- Как вероятна каждая последовательность?
  - P (w1, w2, w3,.. wn)
  - P(w5| w1, w2, w3, w4)

- Языковая модель – математическая модель, которая вычисляется вероятность последовательности слов или условную вероятность следования слова в контексте

Применения

- Распознавание речи
- Статистический машинный перевод
- Исправление опечаток
- Распознавание текстов

- Марковское предположение:
  - На появление слова влияет ограниченное количество предшествующих слов

N-граммы – последовательности n-слов

- Униграммы
- Биграммы
- Триграммы

- Как выбрать

Надежность vs. Точность предсказания

“large green _____”

tree? mountain? frog? car?

“swallowed the large green _____”

pill? broccoli?

Надежность vs. Точность предсказания

- Больше n: больше информации о контексте специфического примера (больше потенциальная точность предсказания)
- Меньше n: больше данных, лучше статистические оценки, больше надежности

Вероятность появления следующего слова

$P(W_n|W_1,.., W_{n-1}) = P(W_1,.....W_n)/P(W_1,...W_{n-1})$

MLE (максимальное правдоподобие)

- $P_{mle}(W_1,...W_n) = C(W_1,....,W_n)/N$
- $P_{mle}(W_n|W_1,.., W_{n-1}) = C(W_1,...W_n)/ C(W_1,...W_{n-1})$
- C ()- частота появления подстроки

Для биграмм

- $Pmle(W_n|W_{n-1}) = C(W_{n-1}, W_n)/C(W_{n-1})$

Пример

- Корпус
  - <s> Он пошел в школу </s>
  - <s> Пошел он в школу</s>
  - <s> Он не любит мясо</s>

- Вероятности по максимальному праводоподобию?:
  - Униграммы: он, пошел, мясо
  - Биграммы: P(он|пошел), P(пошел|он)

Порождая Шекспира

- Порождение предложений из униграмм...
  - Every enter now severally so, let
  - Hill he late speaks; or! a more to leg less first you enter
- С биграммами...
  - What means, sir. I confess she? then all sorts, he is trim, captain.
  - Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry.
- Триграммы
  - Sweet prince, Falstaff shall die.
  - This shall forbid it should be branded, if renown made it empty.

- Тетраграммы
  - What! I will go seek the traitor Gloucester.
  - Will you not tell me who I am?

  - Это выглядит как Шекспир, посколько это и есть Шекспир

- Т.е. как только мы увеличиваем величину N, точность модели растет, так как выбор следующего слова все более ограничен

Проблема

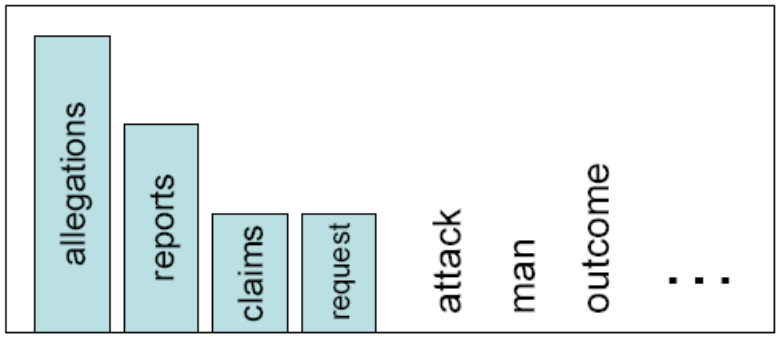
- В тексте, который мы пытаемся предсказать, могут встретиться слова, которых не было

- Методы сглаживания (smoothing)

Smoothing is like Robin Hood: Steal from the rich and give to the poor (in probability mass)

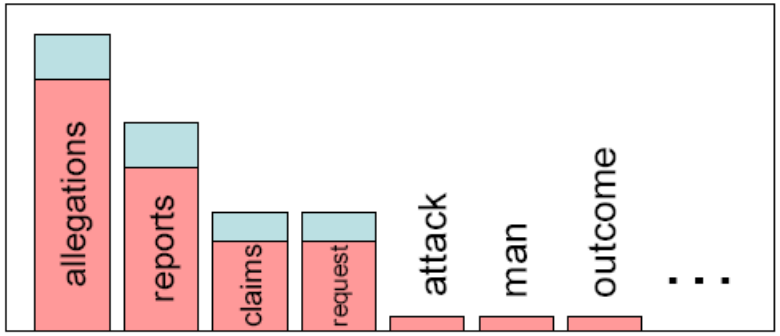
- We often want to make predictions from sparse statistics:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total



- Smoothing flattens spiky distributions so they generalize better

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



- Very important all over NLP, but easy to do badly!

Slide from Dan Klein

Сглаживание Лапласа

- Для униграмм:
  - Добавляем 1 к частоте каждого слова
  - Нормализуем N (#tokens) + V (#types)
  - Исходная вероятность униграммы

$$P(w_i) = \frac{c_i}{N}$$

  - Новая вероятность униграммы

$$P_{LP}(w_i) = \frac{c_i + 1}{N + V}$$

  - Для биграмм
  - исходная

$$P(w_n | w_{n-1}) = \frac{c(w_{n,} w_{n-1})}{c(w_{n-1})}$$

  - новая

$$P(w_n | w_{n-1}) = \frac{C(w_{n-1}, w_n) + 1}{C(w_{n-1}) + V}$$

Статистические оценки

Пример:

Корпус: пять романов Джен Остин

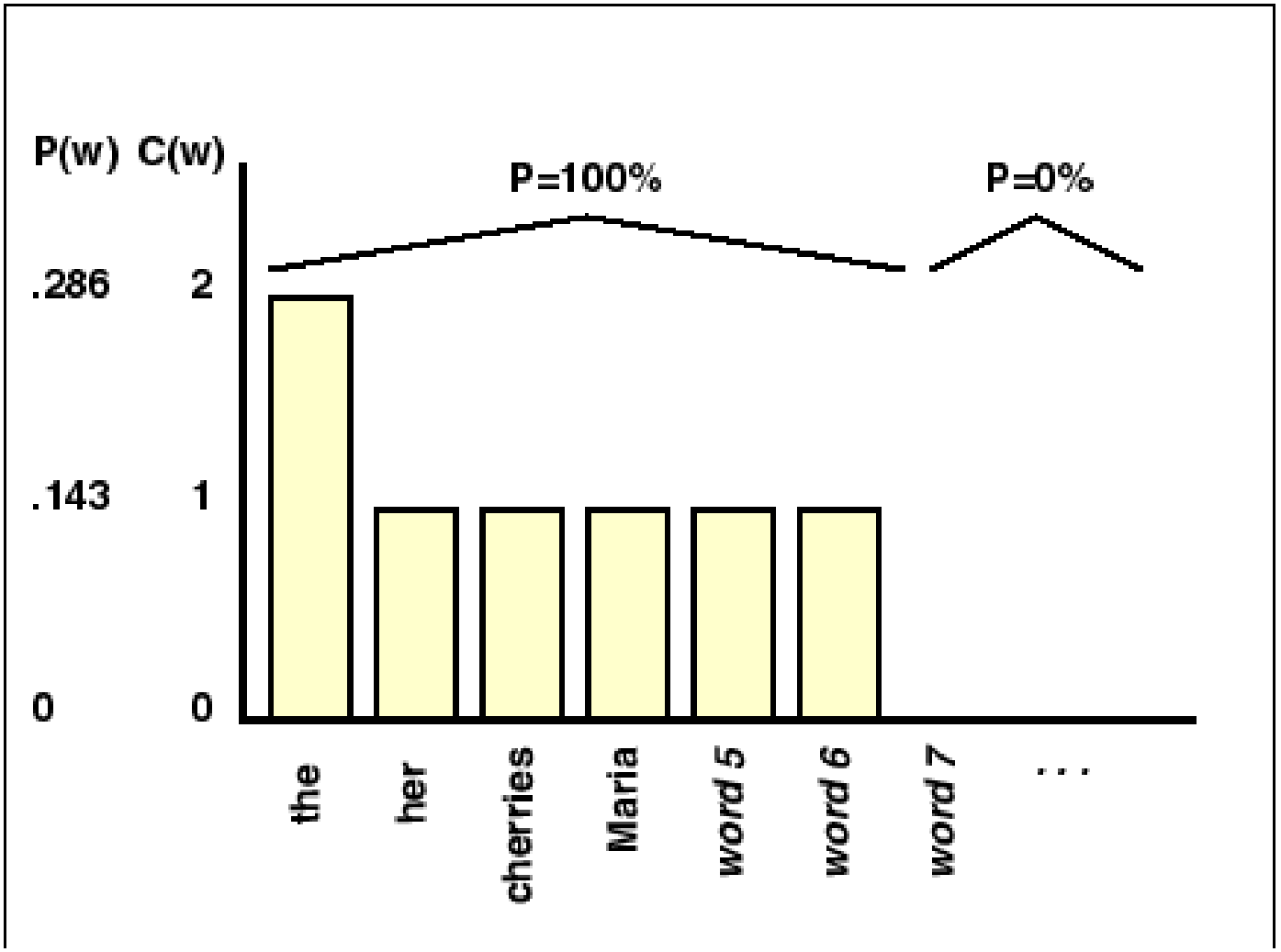
N = 617,091 слов

V = 14,585 уникальных слов

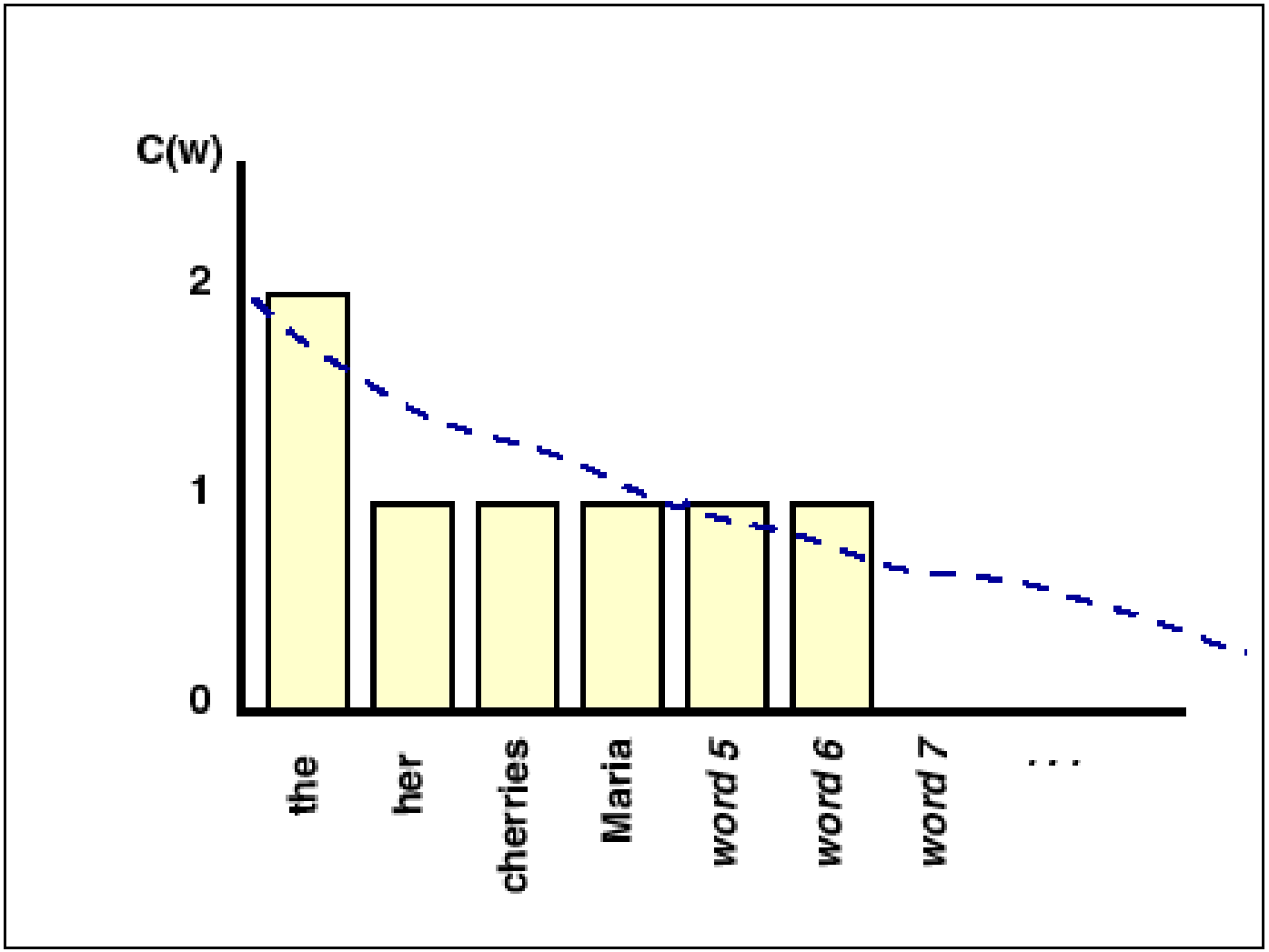
Задача: предсказать следующее слово триграммы “inferior to _____”

from test data, *Persuasion*: “[In person, she was] inferior to *both* [sisters.]”

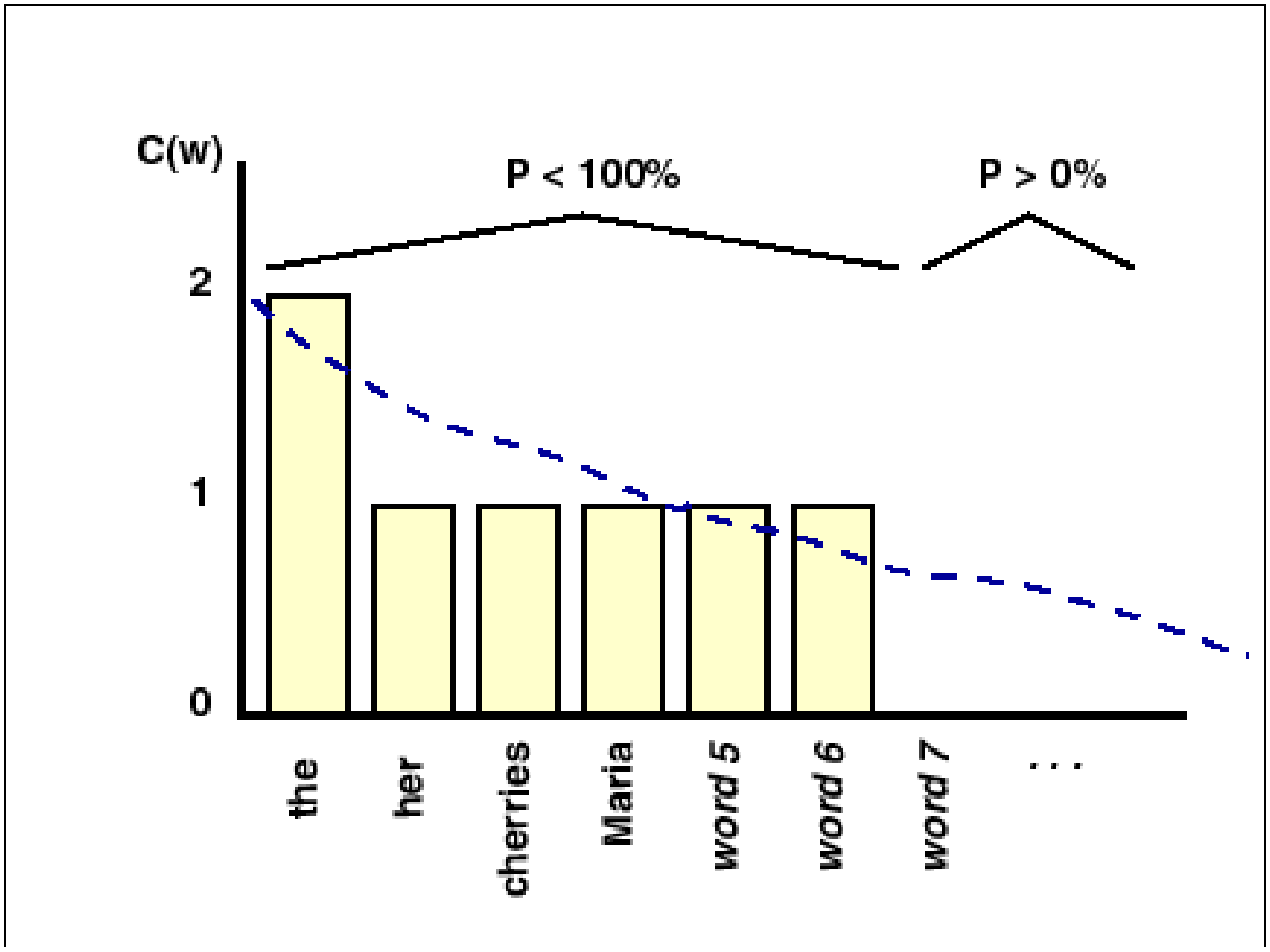
Частотные оценки (MLE):



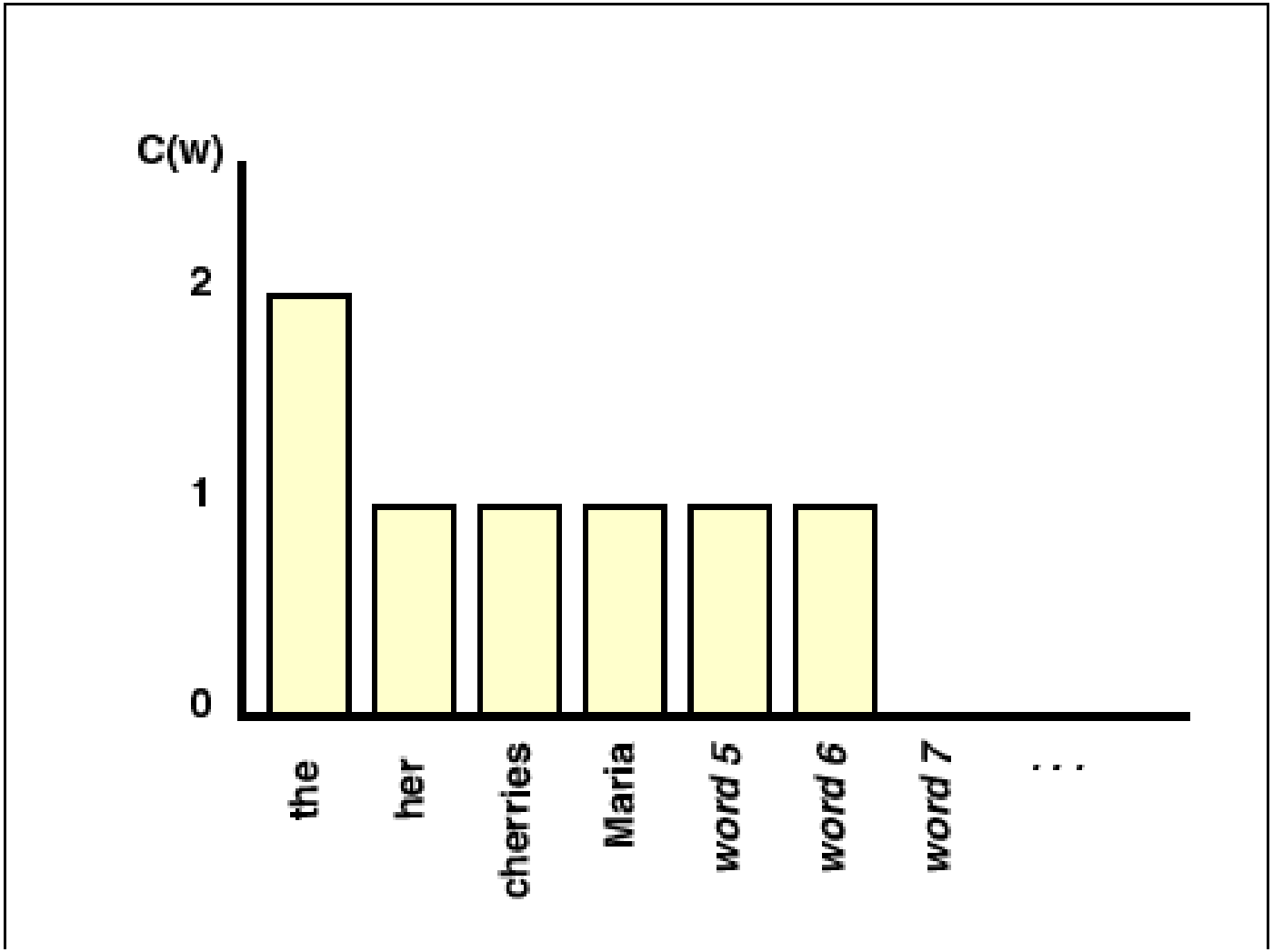
Реальное распределение вероятностей



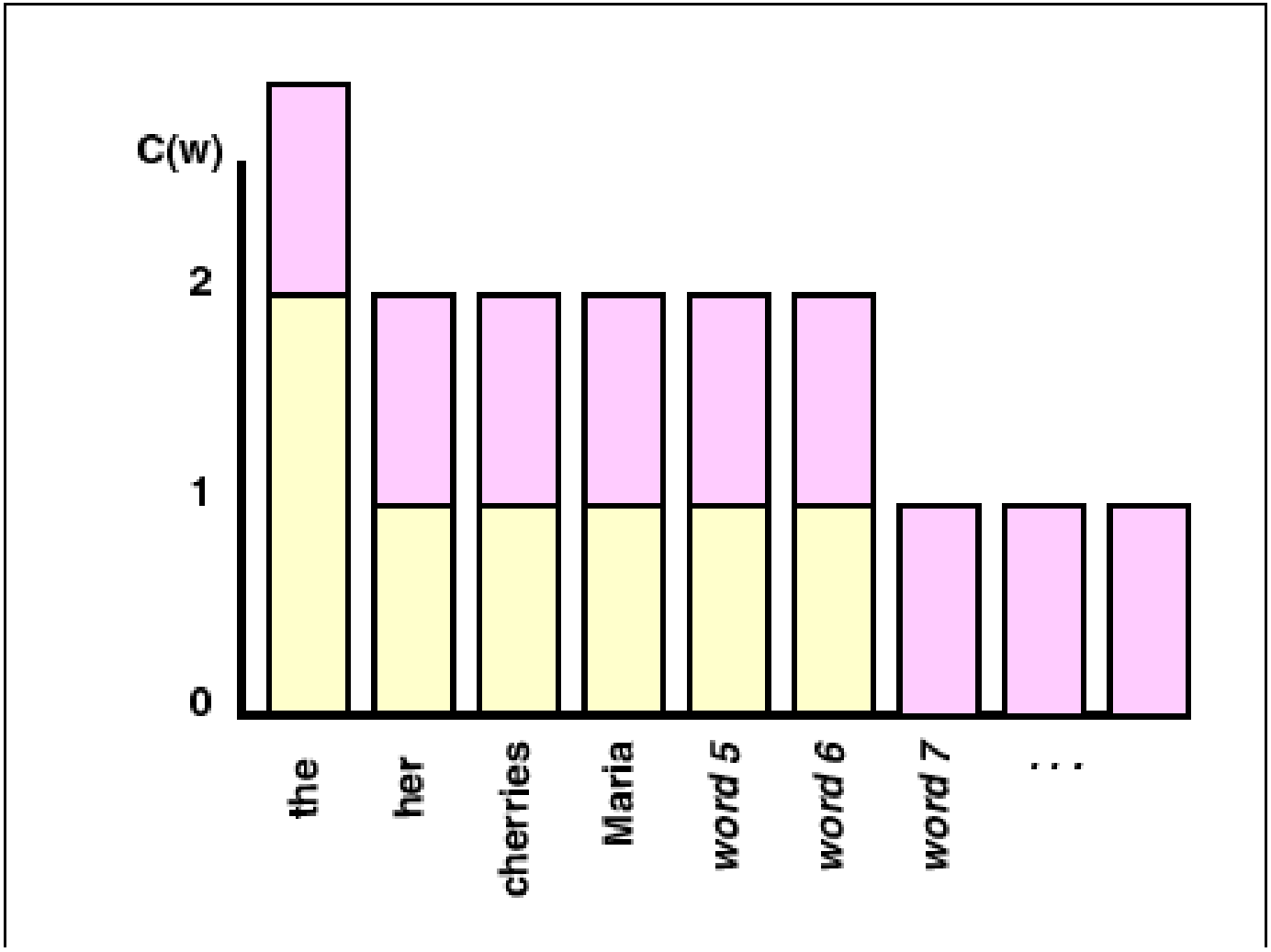
Реальное распределение вероятностей



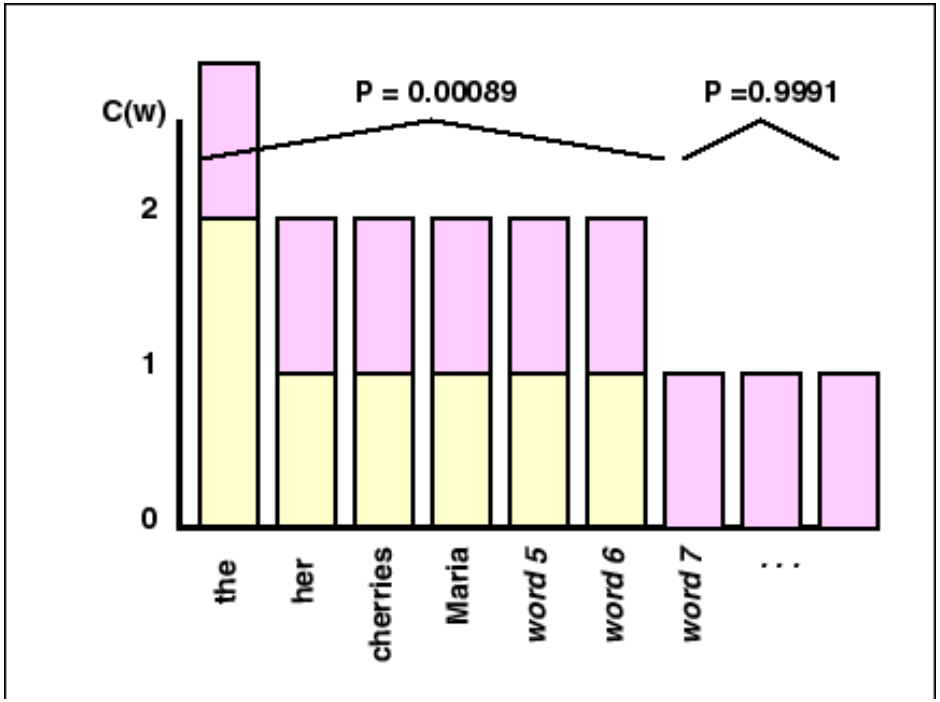
LaPlace's Law
(adding one)



LaPlace's Law
(adding one)



Закон Лапласа



- Более мягкое сглаживание

$$P(w_i) = \frac{C(w_i)+\alpha}{N+\alpha \cdot V}$$

$$P(w_n | w_{n-1}) = \frac{C(w_{n-1}, w_n)+\alpha}{C(w_{n-1})+\alpha \cdot V}$$

Пример

- Корпус
  - <s> Он пошел в школу </s>
  - <s> Пошел он в школу</s>
  - <s> Он не любит мясо</s>

- Вероятности по Лапласу?:
  - Униграммы: он, пошел, мясо
  - Биграммы: P(он|пошел), P(пошел|он)

Однако

- Для больших словарей – биграмм всегда много – закон Лапласа дает слишком много вероятности не встречавшимся событиям

- (Church, Gale, 1991)
- Corpus Associated Press – 44 млн. слов – разделили на две части и пытались предсказать поведение на второй части
- - 400653 разных слов
- - $1.6*10^{11}$ число возможных биграмм

Предсказание числа биграмм

- F_{lap}= ((r+1)/(N+B))*N
- Биграммы встречались r – раз в одной половине корпуса
- Нужно предсказать, сколько раз такие биграммы встретятся во второй половине корпуса

- N=22 млн. слов
- V= 273266 разных слов
- B=V*V

Закон Лапласа и реальные частоты

R=fmle	flap	femp
0	0.000137	0.000027
1	0.000274	0.448
2	0.000411	1.25
3	0.000548	2.24
4	0.000685	3.23
5	0.000822	4.21
6	0.000959	5.23

Flap – предсказание средней частоты во второй части по Лапласу

Femp – реальная средняя частота во второй части

Недостатки!! Метод негибкий, использует очень мало информации об обучающем корпусе

Еще идея для сглаживания

- Использовать $p(t_n|t_2 \ldots t_{n-1})$ для вычисления $p(t_n|t_1 \ldots t_{n-1})$, если $C(t_n|t_1..t_{n-1})=0$

- Модели
  - Откат
  - Интерполяционная модель

Katz's Backing-Off (Модель отката)

- Используем *n-gram* вероятность, когда достаточно данных
  - (когда частота > k; k = {0, 1})
- Если нет, то переходим ("back-off") на *(n-1)-граммную* вероятность
- (Повторяем при необходимости)
- Модель отката и интерполяционная модель оперируют одними и теми же вероятностями
  - но интерполяционная их суммирует
  - модель отката просто переходит на меньшие энграммы, если не хватает информации

Простая линейная интерполяция
(a.k.a., finite mixture models;
a.k.a., deleted interpolation)

$$\boldsymbol{P_{li}(w_n \mid w_{n-2}, w_{n-1}) =}$$

$$\lambda_1 P_1(w_n) + \lambda_2 P_2(w_n \mid w_{n-1}) + \lambda_3 P_3(w_n \mid w_{n-2}, w_{n-1})$$

- Взвешенное среднее униграмм, биграмм и триграмм
- Сумма лямбда =1
-

Пример

- Корпус
  - <s> Он пошел в школу </s>
  - <s> Пошел он в школу</s>
  - <s> Он не любит мясо</s>

- Пусть используется линейная комбинация униграмм, биграмм и триграмм: лямбда=1/3
- Какова вероятность P(в|он, пошел)

Подбор параметров

- Hold out ~ 5 – 10% для тестирования
- Hold out ~ 10% для подбора параметров (smoothing)
- Для тестирования: полезно тестировать на разных коллекциях, и исследовать поведение моделей
- Параметры может быть трудно подбирать.

Рекурсивная интерполяционная модель

$$p_{interp}(w_i|w_{i-n+1}^{i-1}) =$$
$$\lambda_{w_{i-n+1}^{i-1}} p_{ML}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda_{w_{i-n+1}^{i-1}}) p_{interp}(w_i|w_{i-n+2}^{i-1})$$

- Модель n-го порядка определяется рекурсивно через модели низких порядков
- p_n () – это корпусная вероятность
- λ - коэффициент, зависящий от $w_1 \ldots w_{n-1}$
- Чем больше λ, тем больше мы доверяем истории $w_{1,n-1}$
- Много случайных продолжений $w_{1,n-1}$: λ мало
- Продолжений мало и они частотные: $\lambda \approx 1$
- Как определить такие λ?

Для биграмм

- $P_{interp}(w_2|w_1)=\lambda p_{ML}(w_2|w_1)+(1-\lambda)\, p_{interp}(w_2)$

Рекурсивная интерполяционная модель

- Разнообразие предсказанных слов.
- Рассмотрим в корпусе Europarl биграммные истории слов “spite” и “constant”
  - Одинаковая частотность в корпусе 993
  - Только 9 разных слов после “spite”, почти всегда стоит “of ” (979 раз)
  - 415 различных слов после “constant:
    - and (45), concern (27), pressure (26)…
- Это означает, что более вероятно увидеть новую слово (новую биграмму) после слова “constant”

Witten-Bell smoothing

- Интуиция: вероятность увидеть энграмму с нулевой вероятностью свяжем с вероятность увидеть энграмму первый раз:
- $w_1..w_2 \ldots w_1 \ldots w_2 \ldots w_3..$

- $w_1 \ldots$ new w2… $w_1 \ldots w_2$ ..new w_3..

Witten-Bell smoothing

- Метод рекурсивной интерполяции
- Смотрим, сколько возможных продолжений в обучающих данных
- 1- λ - это вероятность необходимости использовать переход на энграммы более низкого порядка.
- Рассмотрим, сколько разных слов встречалось после нграммы $w_1 \ldots w_{n-1}$

$$N_{1+}(w_1, ..., w_{n-1}, \bullet) = |\{w_n : c(w_1, ..., w_{n-1}, w_n) > 0\}|$$

$$1 - \lambda_{w_1,...,w_{n-1}} = \frac{N_{1+}(w_1, ..., w_{n-1}, \bullet)}{N_{1+}(w_1, ..., w_{n-1}, \bullet) + \sum_{w_n} c(w_1, ..., w_{n-1}, w_n)}$$

Результат для spite и constant

$$1-\lambda_{spite} = \frac{N_{1+}(\text{spite},\bullet)}{N_{1+}(\text{spite},\bullet)+\sum_{w_n} c(\text{spite},w_n)}$$

$$= \frac{9}{9+993} = 0.00898$$

$$1-\lambda_{constant} = \frac{N_{1+}(\text{constant},\bullet)}{N_{1+}(\text{constant},\bullet)+\sum_{w_n} c(\text{constant},w_n)}$$

$$= \frac{415}{415+993} = 0.29474$$

Проблема сглаживания Виттена-Белла

- Слово York – м.б. достаточно частотное в коллекции
- Поэтому сглаживание по униграммным вероятностям будет предполагать его достаточно частую встречаемость после разных слов
- Но оно встречается только после слова New
- Проблема: переход на униграммные вероятности дает преувеличенные оценки появления слова в биграмме

Сглаживание Кнейзера Нея (Kneser-Ney Smoothing)

- Рассматривает историю появления слова
- Определим: количество разных энграмм, в которых участвует данное слово

$$N_{1+}(\bullet\ w_i) = |\{w_{i-1} : c(w_{i-1}w_i) > 0\}|$$

$$N_{1+}(\bullet\ \bullet) = \sum_{w_i} N_{1+}(\bullet\ w_i)$$

- Тогда для модели низкого порядка

$$p_{KN}(w_i) = \frac{N_{1+}(\bullet\ w_i)}{N_{1+}(\bullet\ \bullet)}$$

Т.е. вероятность слова определяется не его частотной во всем объеме коллекции, А разнообразием словосочетаний, в которые оно входит

Сглаживание Кнейзера Нея (Kneser-Ney Smoothing)

- Рассматривает историю появления слова
- Униграммная вероятность в биграммной модели считается так:
- - число разных биграмм, где второе слово *wi* по отношению ко всем биграммам в корпусе

$$\hat{p}(w_i) = \frac{N_{1+}(\bullet w_i)}{N_{1+}(\bullet\bullet)}$$

- Ранее униграммная вероятность была

- $P(w_i)=C(w_i)/N$

Как оценить качество языковой модели?

- Обучение на большей части коллекции (Train)
- Проверка на меньшей части (Test)

- Найти вероятность появления коллекции Test – слишком маленькая величина
- Перплексия (Perplexity)

Перплексия

- Перплексия:

$$PP(W) = P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

- Перплексия – это среднее число вариантов, из которых происходит выбор на каждом шаге.
  - Минимизация перплексии – это максимизация вероятности текста
  - Перплексия для предложения, состоящего из случайной последовательности цифр=10

Перплексия для униграмм и биграмм

Униграммы:

$$PP(W) = P(w_1 w_2 ... w_N) = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i)}}$$

Биграммы:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

Расчеты:

$$PP = 2^{-\frac{1}{N}\sum_{1}^{N} \log_2 p(x)}$$

Оценка качества языковой модели
Корпус EuroParl

Перплексия - чем меньше, тем лучше

Smoothing method	bigram	trigram	4-gram
Good-Turing	96.2	62.9	59.9
Witten-Bell	97.1	63.8	60.4
Modified Kneser-Ney	95.4	61.6	58.6
Interpolated Modified Kneser-Ney	94.5	59.3	54.0

Проблемы энграммных моделей

- Много нулевых вероятностей
  - Решение: группирование слов
    - Заранее созданные классы слов
    - Группирование по одинаковым контекстам встречаемости
- Не учитываются дистантные и разрывные связи
  - В большом зеленом доме

Заключение

- Языковые модели присваивают более высокую вероятность предложениям, которые реально встречаются в языке.
- Они используются как компоненты многих систем автоматической обработки текстов, включая распознавание речи и статистический машинный перевод.
- Частотный подход (MLE) дает неточные оценки параметров для разреженного корпуса.
- Техники сглаживания настраивают параметры для оценки вероятности слов, которые не встречались в обучающем корпусе
- В настоящее время новые подходы – на основе нейронных сетей