Оценивание вероятностей появления строк в естественном языке*

Е. А. Будников
unicorn1992@bk.ru
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В работе рассматривается задача оценивания вероятностей появления строк в естественном языке. Для решения задачи используется модель n-грамм. Для решения проблемы большого числа параметров предлагается использовать модель n-грамм на классах. Для решения проблемы нулевых вероятностей строк предлагается использовать три дисконтные модели: Гуда-Тьюринга, Катца и абсолютного дисконтирования.
Вводятся основные определения и описываются методы, а также алгоритм построения классов в модели n-грамм на классах. Описывается проведённый эксперимент на синтетических данных.

Ключевые слова: *языковая модель, дисконтная модель, n-граммы на классах, Гуд-Тьюринг, Катц, абсолютное дисконтирование.*

The estimation of probabilities of appearance of word strings in a natural language*

Y. A. Budnikov
Moscow Institute of Physics and Technology

This article considers the issue of the estimation of probabilities of appearance of word strings in a natural language. N-gram language models are used for solving this issue. Class-based language models are used for solving the problem of huge amount of parameters. Good- Turing estimates, Katz smoothing and absolute discounting smoothing are used for solving the problem of «unseen» words. Basic definitions are introduced the methods and the algorithm of constructing of the classes in class-based language models are described. The work is illustrated by the experiments in the synthetic data.

Keywords: *language model, N-gram language model, class-based language model, Good- Turing estimates, Katz smoothing, absolute discounting smoothing.*

Введение

В задачах, связанных с распознаванием речи, часто возникает необходимость оценивать априорную вероятность появления тех или иных строк. Метод n-грамм описывается в [1, 2, 3, 4] и заключается в том, что апостеорная вероятность появления слова после некой строки зависит не от всех слов строки, а лишь от последних $n-1$.

Основными недостатками этого метода является плохая обучаемость огромного числа параметров и нулевая оценка вероятности появления на строках, которые не встречаются в процессе обучения. Для решения первой проблемы и частичного решения второй предлагается использовать метод n-грамм на классах. Он подробно описывается в [5, 1]. Этот метод заключается в том, что все слова языка разбиваются на классы, тем самым снижается число параметров, затем во время обучения настраиваются вероятности появления в языке шаблонов строк, состоящих из названий классов, а также вероятности появления слова в определённом классе.

Количество строк с нулевой вероятностью уменьшается, однако они остаются. Для перераспределения вероятностей предлагается использовать различные дисконтные

Научные руководители: В. В. Стрижов, В. Я. Чучупал

модели[3, 1, 4]. В модели Гуда-Тьюринга[6] все n-граммы разбиваются на группы в зависимости от частоты появления, а затем происходит сглаживание этих частот между соседними группами. Этот метод прост в реализации, однако неустойчив. Что означает эта неустойчивость, будет пояснено ниже. Также он сглаживает и оценки вероятностей n-грамм, которые встречаются в обучении достаточно часто и могут быть признаны надёжно обученными.

В модели Катца[7] выбирается соответствующий порог, и оценки вероятностей n-грамм, частота появления которых в обучении больше этого порого, не сглаживаются. Однако эта модель также неустойчива.

Модель абсолютного дисконтирования[8] использует другой подход. Из всех ненулевых частот вычитается фиксированное число, которое потом перераспределяется между n-граммами, не встретившимися в обучении. Можно подобрать это число так, чтобы суммарное уменьшение вероятности было таким же, как и в модели Гуда-Тьюринга.

Все эти методы были описаны в обзоре[9].

В данной работе предложены и реализованы несколько комбинаций алгоритмов оценивания вероятностей появления строк в естественном языке.

Постановка задачи

Пусть $W = \overline{w_1 w_2 \dots w_k}$ — строка из слов w_i, принадлижащих словарю Ω, которую подают на вход зашумлённого канала. Роль такого канала могут исполнять радиоэфир или человек, который переводит строку на другой язык. На выходе получим сигнал Y. По этому сигналу необходимо восстановить исходную строку. Чтобы минимизировать вероятность ошибки, необходимо взять такую строку $\hat{W}$, апостеорная вероятность которой $\Pr(\hat{W}|Y)$ максимальна:

$$\hat{W} = \arg\max_{W \in \Omega^*} \Pr(W|Y). \tag{1}$$

При фиксированном выходе Y эта задача эквивалентна максимизации совместной плотности строки W и выхода Y $\Pr(W, Y)$. Но при этом по формуле Байеса получим:

$$\Pr(W, Y) = \Pr(Y|W) \cdot \Pr(W). \tag{2}$$

Получили разбиение большой задачи на две подзадачи. Данная работа посвящена оцениванию второго множителя $\Pr(W)$.

Описание моделей

Будем обозначать подстроку строки W $w_i^j = \overline{w_i w_{i+1} \dots w_j}$, где i — позиция первого символа подстроки, а j — позиция последнего. При таких обозначениях $W \equiv w_1^k$. По формуле Байеса вероятность появления строки раскладывается в произведение апостеорных вероятностей появления каждого слова этой строки при условии извесной «предыстории», то есть подстроки, предшествующей данному слову:

$$\Pr(w_1^k) = \Pr(w_k|w_1^{k-1}) \cdot \Pr(w_{k-1}|w_1^{k-2}) \cdot \ \dots \ \cdot \Pr(w_2|w_1) \cdot \Pr(w_1) \tag{3}$$

В [9] было введено определение модели естественного языка:

Определение 1. *Моделью естественного языка назовём семейство функций*

$$f : \mathbb{R}^P \times \mathbb{R}^N \to \mathbb{R}^k,$$

где $\mathbb{R}^P$ — пространство параметров, $\mathbb{R}^N$ — пространство акустических входов, $\mathbb{R}^k$ — пространство прогнозов

Существует также и другое определение, но уже статистической модели языка[2].

Определение 2. *Статистической моделью естественного языка семейство функций*

$$f : \mathbb{R}^P \times \Omega^* \to [0, 1],$$

где $\mathbb{R}^P$ — пространство параметров, Ω^* — пространство строк, составленных из слов словаря Ω, [0, 1] — оценка вероятности появления строки в языке

Самым распространённым критерием качества модели является уровень ошибок прогнозирования. Однако измерение этого уровня требует участия систем распознавания речи. Однако можно оценивать качество модели и без их участия по тестовым строкам текста. Качество оценивается величиной *перплексии*[2].

Определение 3. *Перплексией назовём следующую величину:*

$$PP = \Pr(w_1 w_2 \dots w_k)^{-\frac{1}{k}}.$$

Перплексия является величиной, обратной к величине средней вероятности, приписываемой каждому слову тестовой строки. Модель обладает большей перплексией, если число слов, которые могут идти после заданного предыдущего, в среднем больше. Таким образом, перплексия является мерой сложности модели.

Модель n-грамм

Если не вводить никаких предположений по поводу вероятностей вида $\Pr(w_k|w_1^{k-1})$, то число параметров будет равно числу всевозможных строк языка, то есть бесконечным растущим с ростом длины строки.

В методе n-грам мы считаем две предыстории одинаковыми, если они оканчиваются на одинаковые $n-1$ слов. Другими словами,

Определение 4. *Модель естественного языка называется моделью на n-граммах, если для параметров модели выполнено условие:*

$$\Pr(w_k|w_1^{k-1}) = \Pr(w_k|w_{k-n+1}^{k-1}). \tag{4}$$

Пример. Статистическая модель биграмм задаёт следующее семейство функций:

$$f = \Pr(w_1 w_2 \dots w_n) = \Pr(w_n|w_{n-1}) \cdot \Pr(w_{n-1}|w_{n-2}) \cdot \ \dots \ \cdot \Pr(w_2|w_1) \cdot \Pr(w_1)$$

Относительно числа параметров такой модели имеет место слеследующая

Лемма 1. *Если словарь содержит V слов, то модель n-грамм содержит $V^n - 1$ параметров.*

Если словарь содержит V слов, то 1-граммы (или *униграммы*) порождают модель, имеющую $V-1$ независимых параметров: V параметров $\Pr(w_i)$ связаны равенством

$$\sum_{i=1}^{V} \Pr(\tilde{w}_i) = 1, \tag{5}$$

где $\tilde{w}_i$ — слова из словаря. 2-граммы (или *биграммы*) порождают $V^2 - 1$ независимых параметров: $V(V-1)$, имеющих форму $\Pr(w_2|w_1)$, и $V-1$, имеющих форму $\Pr(w)$. Далее по индукции легко показать, что модель n-грамм содержит $V^n - 1$ параметров.

Действительно, $V^{n-1}(V-1)$ параметров, имеющих форму $\Pr(w_n|w_1^{n-1})$, и $V^{n-1} - 1$ параметров более низкого порядка(по предположению индукции). Всего

$$V^{n-1}(V-1) + V^{n-1} - 1 = V^n - 1.$$

Настраивать параметры модели будем по тексту T.

Пусть $C(\mathbf{w})$ — число раз, которые строка $\mathbf{w}$ встретилась в обучающем тексте. Тогда в случае *униграмм* максимум правдоподобия для параметра $\Pr(w)$ достигается при $\Pr(w) = \frac{C(w)}{T}$.

Для случая n-грамм имеет место такой результат максимизации правдоподобия:

$$\Pr(w_n|w_1^{n-1}) = \frac{C(w_1^{n-1}w_n)}{\sum_w C(w_1^{n-1}w)}. \tag{6}$$

Модель n-грамм на классах

Для улучшения надёжности обучения параметров необходимо уменьшать их число, стараясь при этом не сильно потерять в точности оценок вероятностей. Также существует проблема нулевых оценок вероятностей появления строк в языке. Приведём пример. Допустим в обучающей выборке текстов многократно и в похожих ситуациях употребляются слова «мяч» и «мячик», за одним исключением: сочетания «уронила в речку мячик» и «не утонет в речке мяч» в нём присутствуют, а «уронила в речку мяч» и «не утонет в речке мячик» в нём отсутствуют. Получится, что оценка вероятностей появления соответствующих строк не только кардинально отличаются, но и пара из этих оценок и вовсе оказываются нулевыми, хотя интуитивно оценки для соответстствующих пар строк практически не должны отличаться.

Эти общие соображения естественным образом подводят нас к идее классов.

Пусть существует некоторая функция $\pi : \Omega \to G$, где Ω — множество слов, словарь, а G — множество классов слов. Тогда обозначим $Pr(w|g)$ вероятность появления в языке слова w, если известен его класс g, а $\Pr(g_n|g_1^{n-1})$ — вероятность встретить слово из класса g_n после последовательности слов, имеющих форму $g_1 g_2 \dots g_{n-1}$.

Теперь мы пожертвуем частью информации, а именно, будем настраивать лишь параметры вида $\Pr(g_n|g_1^{n-1})$ и $Pr(w|g)$.

Определение 5. *Модель n-грамм назовём моделью n-грамм на классах, если выполняется гипотеза:* $\Pr(w_k|w_1^{k-1}) = \Pr(w_k|g)\Pr(g_k|g_1^{k-1})$,*где* $k = 1, \dots, n$.

Пример. Статистическая модель биграмм на классах задаёт следующее семейство функций:

$$f = \Pr(w_1 w_2 \dots w_n) = \Pr(w_n|g_n) \cdot \Pr(g_n|g_{n-1}) \cdot \ \dots \ \cdot \Pr(w_2|g_2) \cdot \Pr(g_2|g_1) \cdot \Pr(w_1|g_1) \cdot \Pr(g_1)$$

Относительно числа параметров такой модели имеет место слеследующая

Лемма 2. *Если словарь содержит V слов и имеется C классов, то модель n-грамм на классах содержит $C^n + V - C - 1$ параметров.*

Действительно, имеется $C^n - 1$ параметров вида $\Pr(g_n|g_1^{n-1})$ (доказывается аналогично Лемме 1) и $V - C$ параметров вида $Pr(w|g)$, так всего таких вероятностей V, но для каждого класса $g \in G$ выполняется равенство:

$$\sum_{w:\pi(w)=g} \Pr(w|g) = 1. \tag{7}$$

Опишем теперь один алгоритм построения функции π на примере биграмм.

Пусть $T = (t_1, t_2, \ \dots \ , t_T)$ — обучающая выборка, причём все слова содержатся в словаре Ω. Функция правдоподобия тогда равна

$$L(T) = \Pr(T) = \prod_{x,y \in \Omega} \Pr(y|x)^{C(xy)}, \tag{8}$$

где x,y — слова из словаря, а $C(xy)$ показывает, сколько раз последовательность слов «xy» встретилась в обучающей выборке T.

Решается максимизационная задача:

$$L(T) \to \max_{\pi}. \tag{9}$$

Покажем, что имеет место

Лемма 3. *Задача максимизации 9 равносильна максимизации функции*

$$F_\pi = \sum_{g,h\in G} C(gh)\cdot \log C(gh) - 2\sum_{h\in G} C(h)\cdot \log C(h),$$

где $C(gh)$ — функция, которая показывает, сколько раз в обучающем тексте встретились строки вида «xy», где $\pi(x) = g$, а $\pi(y) = h$

Для удобства будем использовать логарифм функции правдоподобия вместо самой функции:

$$\log L(T) = \sum_{x,y\in\Omega} C(xy)\cdot \log \Pr(y|x). \tag{10}$$

Из данного выше определения модели n-грамм на классах заключаем, что максимум правдоподобия для биграмм достигается при

$$\Pr(w_i|w_{i-1}) = \frac{C(w_i)}{C(\pi(w_i))}\cdot\frac{C(\pi(w_{i-1})\pi(w_i))}{C(\pi(w_{i-1}))}, \tag{11}$$

где $C(w_i)$ — число раз, которые слово w_i встретилось в обучающей выборке, а $C(\pi(w))$ — число раз, которые слова из класса $\pi(w)$ встретились в выборке, аналогично $C(\pi(w_x)\pi(w_y))$ — число пар вида «$\pi(w_x)\pi(w_y)$», встретившиеся в выборке.

Подставим теперь это выражение в функцию правдоподобия и преобразуем:

$$\log L(T) = \sum_{x,y\in\Omega} C(xy)\cdot \log\left(\frac{C(y)}{C(\pi(y))}\cdot\frac{C(\pi(x)\pi(y))}{C(\pi(x))}\right) \tag{12}$$

$$= \sum_{x,y\in\Omega} C(xy)\cdot \log\left(\frac{C(y)}{C(\pi(y))}\right) + \sum_{x,y\in\Omega} C(xy)\cdot \log\left(\frac{C(\pi(x)\pi(y))}{C(\pi(x))}\right)$$

$$= \sum_{y\in\Omega} C(y)\cdot \log\left(\frac{C(y)}{C(\pi(y))}\right) + \sum_{g,h\in G} C(gh)\cdot \log\left(\frac{C(gh)}{C(g)}\right)$$

$$= \sum_{y\in\Omega} C(y)\cdot \log C(y) - \sum_{y\in\Omega} C(y)\cdot \log C(\pi(y))$$

$$+ \sum_{g,h\in G} C(gh)\cdot \log C(gh) - \sum_{g,h\in G} C(gh)\cdot \log C(g)$$

$$= \sum_{y\in\Omega} C(y)\cdot \log C(y) + \sum_{g,h\in G} C(gh)\cdot \log C(gh)$$

$$-2\sum_{h\in G} C(h)\cdot \log C(h).$$

Заметим, что первое слагаемое не зависит от выбора функции π. Поэтому его рассматривать необязательно, когда мы будем оптимизировать π.

Алгоритм 1 Алгоритм построения функции π.

1: **для всех** $w \in \Omega$
2: $G(w) = 1$ //инициализация
3: **для** $i = 1 \dots n$
4: **повторять**
5: **для всех** $c \in G$
6: Переместить слово w в класс c, запомнив его предыдущий класс
7: Вычислить изменения F_π для этого перемещения в c. Переместить слово w назад в его предыдущий класс
8: Переместить слово w в класс, кторый больше всего увеличивает F_π, или никуда не перемещать, если увеличения ни на каком перемещении не происходит
9: **пока** s

Поэтому будем максимизировать функцию

$$F_\pi = \sum_{g,h \in G} C(gh) \cdot \log C(gh) - 2 \sum_{h \in G} C(h) \cdot \log C(h). \tag{13}$$

Приведём теперь алгоритм оптимизации функции π. Перед запуском алгоритма определяется число классов.

Имеет место следующее утверждение:

Лемма 4. *Алгоритм 1 сходится к локальному минимуму F_π.*

Утверждение очевидно и следует из того, что на каждом шаге значение F_π увеличивается.

Дисконтная модель

Рассмотрим событие S, которое встретилось s раз, а общее количество наблюдений A. Тогда оценка вероятности S по принципу наибольшего правдоподобия будет равна

$$\Pr(S) = \frac{s}{A}. \tag{14}$$

Но тогда, в соответствии с этим принципом, событиям, которые не были встречены среди обучающего текста T, будут приписаны нулевые вероятности, а значит, будучи встреченными на тесте, они никогда не будут распознаны.

Чтобы справиться с этой проблемой, можно поступить следующим способом. В оценке вероятности события вместо числа s брать

$$s' = d_s \cdot s, \tag{15}$$

где d_s — множитель, зависящий от числа раз, которые событие встретилось в обучающем тексте. Тогда получим дисконтную оценку вероятности события S:

$$\Pr{}_{discount}(S) = \frac{s'}{A} = \frac{d_s \cdot s}{A}. \tag{16}$$

Различные дисконтные методы различаются стратегией выбора d_s.

Обозначим c_s число всех событий которые встретились в процессе обучения ровно s раз. Тогда общее число наблюдений $A = \sum_{s \geqslant 1} c_s \cdot s$. Получается, что таким образом мы перераспределили оценки вероятности между событиями и оставили на все не встретившиеся в обучении слова $1 - \frac{1}{A} \sum_{s \geqslant 1} d_s \cdot c_s \cdot s$. Если c_0 — число таких событий, то оценка вероятности каждого из них равна

$$\frac{1}{c_0} \left(1 - \frac{1}{A} \sum_{s \geqslant 1} d_s \cdot c_s \cdot s \right). \tag{17}$$

Дисконтная модель Гуда-Тьюринга

В статье[6] предлагается следующая стратегия выбора множителя:

$$d_s = (s+1)\frac{c_{s+1}}{s \cdot c_s}. \tag{18}$$

Эта стратегия называется оценкой Гуда-Тьюринга. Несмотря на очевидную простоту стратегии, у неё есть существенный недостаток: она проваливается в случае, если $c_a = 0$ для некоторого a и существует $b > a$, такой, что $c_b \neq 0$. Также существенно, что дисконтирование необходимо для параметров, оценка которых является ненадёжной, то есть для тех событий, которые встречаются в обучении менее некоторого количества раз k, выбранного априори.

Дисконтная модель Катца

Решение этой проблемы было предложено в [7]. Пусть есть некое, достаточно большое число k, такое что все оценки вероятностей событий, встретившихся в процессе обучения более k раз, признаем надёжными. При этом d_s будет выглядеть так:

$$d_s = \begin{cases} \frac{(s+1)\frac{c_{s+1}}{s \cdot c_s} - (k+1)\frac{c_{k+1}}{c_1}}{1-(k+1)\frac{c_{k+1}}{c_1}} & , \ 1 \leqslant s \leqslant k \\ 1 & , \ s > k \end{cases} \tag{19}$$

Этот метод тоже нестабильный, так как возможны ситуации, когда $d_s < 0$.

Модель абсолютного уменьшения

Одной из альтернатив модели Гуда-Тьюринга является модель абсолютного уменьшения[8]. В этой модели происходит уменьшение числа a для каждого события на фиксированное число m.

$$d_s = \frac{s-m}{s}. \tag{20}$$

Для того чтобы уменьшение суммарной вероятности было таким же, как в модели Гуда-Тьюринга, необходимо, чтобы

$$m = \frac{c_1}{\sum_{s \geqslant 1} c_s}. \tag{21}$$

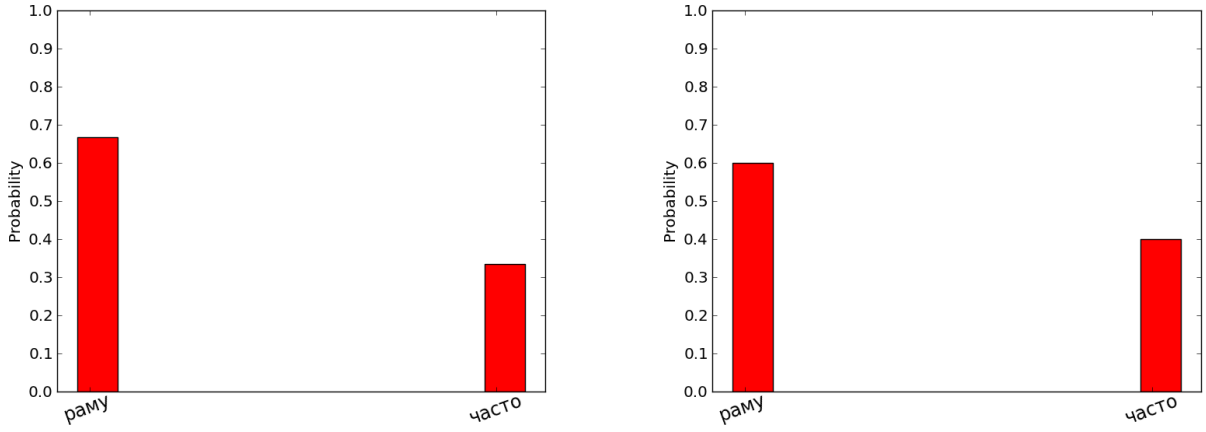
Вычислительный эксперимент

Целью вычислительного эксперимента являлась демонстрация работы комбинаций алгоритмов на небольшом массиве синтетических данных, состоящих из небольшого текста на тему «Мама моет раму». В первой серии экспериментов оценивалось распределение вероятностей появления слова после заданной строки текста. Во второй серии экспериментов оценивалась перплексия различных тестовых строк: встречающейся в обучающем тексте и двух не встречающихся в тексте.

В обеих сериях для методов n-грамм и n-грамм на классах проводилось по четыре типа экспериментов: без дисконтирования и по каждому из трёх типов дисконтирования.

Оценивание распределения вероятностей после заданной фразы

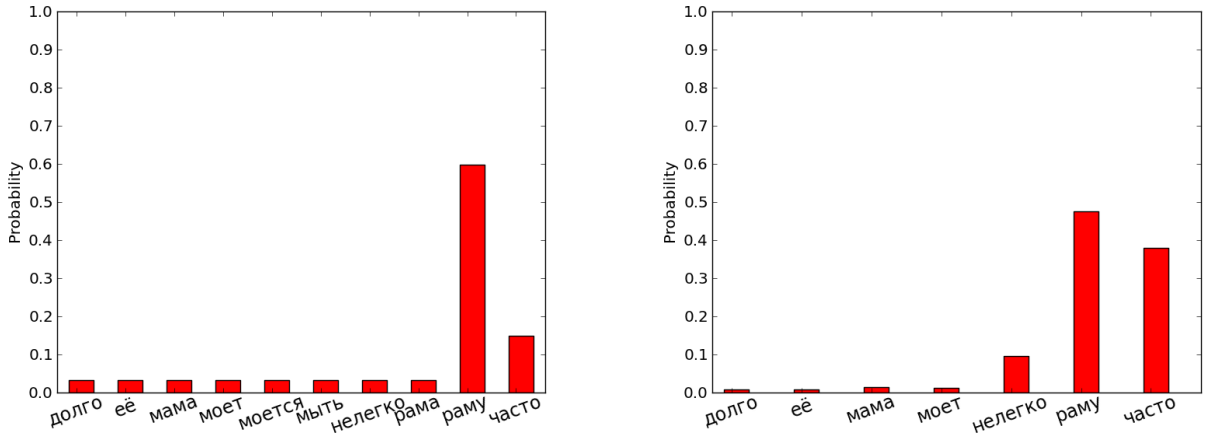
В первой серии экспериментов оценивалось распределение вероятностей появления слова после фразы «Мама моет...»



Метод n-грамм Метод n-грамм на классах

Рис. 1. Методы без дисконтирования

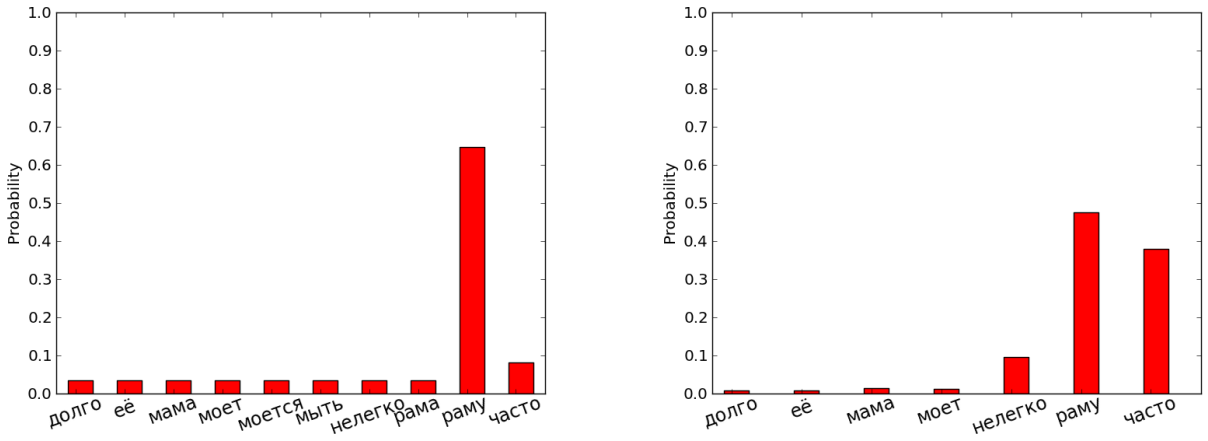
Оба метода без дисконтирования сработали примерно одинаково, распределив лишь немного иначе вероятности между двумя вариантами продолжения. Также читатель может заметить, что метод n-грамм на классах немного сгладил разницу между вероятностями. Это связано с тем, что алгоритм 1 определил слова «раму» и «часто» в один класс, а вероятности между этими словами распределяются в зависимости от суммарной частоты появления в обучающем тексте, а не только после строки «Мама моет...», а точнее, шаблона строки «$g_1 g_2 ...$», где g_1 — класс слова «мама», а g_2 — класс слова «моет».



Метод n-грамм Метод n-грамм на классах

Рис. 2. Модель дисконтирования Гуда-Тьюринга

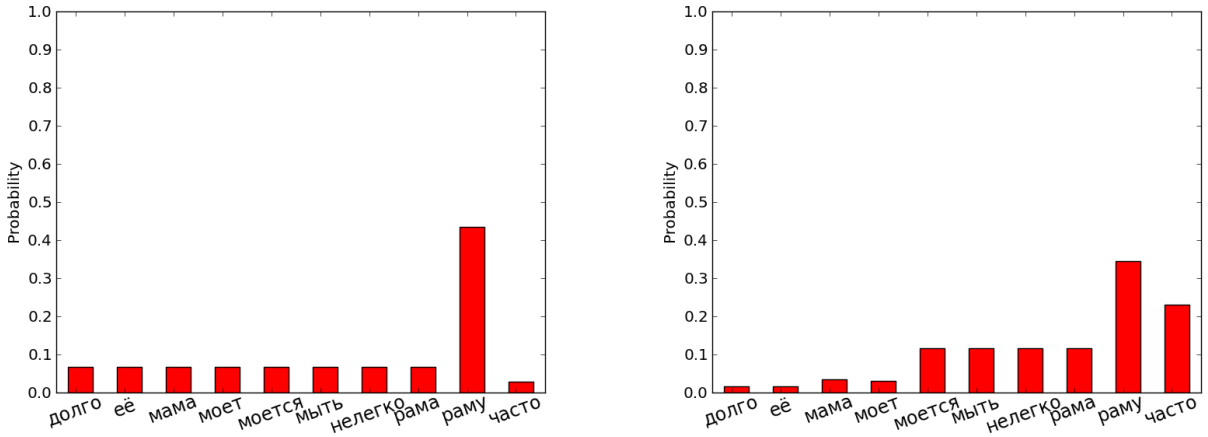
На рисунке 2 продемонстрирован метод дисконтирования Гуда-Тьюринга. В графики были включены лишь варианты с вероятностями > 0.004. Читатель может обратить внимание на снизившуюся оценку вероятности слова «часто» в методе n-грамм. Это связано с тем, что метод дисконтирования предполагает, что оценка вероятности появления события, встретившегося однажды или дважды в процессе обучения, не должна существенно отличаться от оценки вероятности появления события, в процессе обучения не встретившегося.



Метод n-грамм Метод n-грамм на классах

Рис. 3. Модель дисконтирования Катца

На рисунке 3 можно заметить, что в модели дисконтирования Катца надёжно обученные параметры не сглаживаются.



Метод n-грамм Метод n-грамм на классах

Рис. 4. Модель асболютного дисконтирования

В методе абсолютного дисконтирования при использовании метода n-грамм происходит парадоксальная ситуация: оценка вероятности события, встречавшегося в обучении, в итоге оказывается ниже оценки вероятности события, которое в обучении не встретилось.

Это объясняется высокой долей биграмм, которые встретились в обучении только один раз, среди всех встретившихся в обучении биграмм.

Оценка сложности модели на тестовых строках

Во второй серии экспериментов оценивалась перплексия различных тестовых строк.

Таблица 1. Перплексия подстроки из обучающего текста «Мама моет часто».

Модель дисконтирования	n-граммы	n-граммы на классах
Без дисконтирования	2.5	2.06186
Гуд-Тьюринг	5.62562	3.82051
Катц	9.66017	1.9987
Абсолютное	3.00793	2.71695

Таблица 2. Перплексия подстроки «Мама моет долго», которая не встречается в обучающем тексте.

Модель дисконтирования	n-граммы	n-граммы на классах
Без дисконтирования	∞	6.12372
Гуд-Тьюринг	12.2359	25.5717
Катц	14.8572	13.3778
Абсолютное	8.53946	18.0222

Читатель может заметить по таблицам 1 и 2, что самым предпочтительным пока выглядят метод n-грамм на классах без дисконтирования и метод n-грамм на классах с дисконтированием Катца. Они надёжно оценивают вероятности строк и обладают минимальными перплексиями.

Однако давайте посмотрим на оценку перплексии ещё одной строки.

Таблица 3. Перплексия подстроки «Долго её моет».

Модель дисконтирования	n-граммы	n-граммы на классах
Без дисконтирования	∞	∞
Гуд-Тьюринг	6.53827	6.14817
Катц	6.84374	8.76256
Абсолютное	8.12289	7.77689

В таблице 3 читатель может заметить, что если метод n-грамм на классах без дисконтирования даёт нулевую оценку вероятности строки, то более предпочтительными являются методы с дисконтированием Гуда-Тьюринга или абсолютным дисконтированием.

Заключение

В работе были рассмотрены методы оценивания вероятностей появления строк в языке, основанные на n-граммах. Каждый из рассмотренных методов обладает, как показал вычислительный эксперимент, своими достоинствами и недостатками. К достоинствам метода n-грамм без дисконтирования можно отнести линейную по размеру обучающего текста сложность алгоритма настройки параметров, к недостаткам — большое число параметров и, как следствие, плохую их обучаемость, а также нулевую оценку вероятности появления в языке n-грамм, которые не встретились в процессе обучения.

К достоинствам метода n-грамм на классах можно отнести, что число параметров линейно по размеру словаря и квадратично по числу классов, локальную оптимальность решения задачи разбиения слов на классы. Недостатками являются высокая вычислительная сложность алгоритма, а также наличие нулевых оценок вероятностей, хоть и на меньшем количестве строк с сравнении с методом n-грамм.

Дисконтные модели решают проблему нулевых оценок вероятностей появления строки в естественном языке, однако они могут работать неадекватно, если велика доля ненадёж-

но обученных параметров. Также недостатком моделей Гуда-Тьюринга и Катца является их неустойчивость.

Литература

[1] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing, A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, 2001.

[2] Frederick Jelinek. *Statistical Methods for Speech Recognition.* The MIT Press, Cambridge, Massachusetts, 1997.

[3] Yoshihiko Gotoh and Steve Renals. Statistical language modelling. In Steve Renals and Gregory Grefenstette, editors, *ELSNET Summer School*, volume 2705 of *Lecture Notes in Computer Science*, pages 78–105. Springer, 2000.

[4] Steve Young and Gerrit Bloothooft, editors. *Corpus-Based Methods in Language and Speech Processing.* Kluwer Academic Publishers, Dordrecht, 1997.

[5] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, and Robert L. Mercer. Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, pages 283–298, Paris, France, March 1990.

[6] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3 and 4):237–264, 1953.

[7] Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401, March 1987.

[8] Hermann Ney, Ute Essen, and Reinhard Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[9] Егор Алексеевич Будников. Обзор некоторых статистических моделей естественных языков. *Машинное обучение и анализ данных*, 1:245–250, декабрь 2011.