Budnikov E.A., Strijov V.V. Estimating probabilities of text strings in document collections // Information Technologies, 2013, 4 : 40-45.

УДК 519.256

Е. А. Будников, студ. Московский физико-технический институт

В. В. Стрижов, к.ф.-м.н., н.с., Вычислительный центр РАН

Оценивание вероятностей появления строк в коллекции документов[1]

В работе рассматривается задача оценивания вероятностей появления строк в документах. 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 появления строк используется три дисконтные модели: Гуда-Тьюринга, Катца и абсолютного дисконтирования. Описывается проведённый эксперимент на синтетических данных. Предлагаемая модель проиллюстрирована вычислительным экспериментом на реальных данных.

Ключевые слова: *языковая модель, дисконтная модель, n-граммы на классах, модель Гуда-Тьюринга, модель Катца, абсолютное дисконтирование.*

Введение

В задачах, связанных с анализом текста, требуется оценить априорную вероятность появления строк. Для этого используется метод n-грамм [1, 2, 3, 4], который заключается в том, что апостеорная вероятность появления слова после некой строки зависит не от всех слов строки, а лишь от последних $n-1$ слов.

Основными недостатками этого метода являются, во-первых, сложность получения оценок большого числа параметров статистической модели и,

1 Работа выполнена при поддержке Министерства образования и науки РФ в рамках Государственного контракта 07.524.11.4002.

во-вторых, проблема наличия нулевых оценок вероятности появления слов в строках, которые не встречаются в процессе обучения. Для решения этих проблем предлагается использовать метод n-грамм на классах. 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 [5, 1]. Количество строк с нулевой вероятностью уменьшается, однако они остаются.

Для перераспределения вероятностей предлагается использовать различные дисконтные 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 порога, 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. В данной работе предложены и реализованы алгоритмы оценивания вероятностей появления строк в коллекции документов.

1. Постановка задачи

Пусть $W = \overline{w_1 w_2 \dots w_k}$ — строка из слов w_i из словаря Ω и Y –описание,

соответствующее этой строке. По описанию необходимо восстановить исходную строку. Чтобы минимизировать вероятность ошибки, необходимо взять такую строку $\hat{W}$, апостериорная вероятность которой $\Pr(\hat{W} \mid Y)$ максимальна:

$$\hat{W} = \arg\max_{W \in \Omega} \Pr(W \mid Y). \quad (1)$$

При фиксированном описании Y эта задача эквивалентна максимизации совместной плотности $\Pr(W, Y)$ строки W и описания Y. Эта функция может быть представлена в виде произведения:

$$\Pr(W, Y) = \Pr(Y \mid W) \cdot \Pr(W). \quad (2)$$

функции правдоподобия вектора Y и априорной функции вероятности строки W. Данная работа посвящена оцениванию второго множителя $\Pr(W)$.

2. Описание моделей

Обозначим подстроку строки W как $w_i^j = \overline{w_i w_{i+1} \ldots w_j}$, где i — позиция первого символа подстроки, а j — позиция последнего. Согласно этому обозначению строка $W \equiv w_1^k$. Вероятность появления строки равна произведению апостеорных вероятностей появления каждого слова этой строки при условии известной предыстории, то есть подстроки, предшествующей данному слову:

$$\Pr(w_1^k) = \Pr(w_k \mid w_1^{k-1}) \cdot \Pr(w_{k-1} \mid w_1^{k-2}) \cdot \ldots \cdot \Pr(w_2 \mid w_1) \cdot \Pr(w_1). \quad (3)$$

Определение 1. *Моделью естественного языка назовём семейство функций*

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

где R^P — пространство параметров, R^N — пространство исходных строк, R^k — пространство зависимых переменных.

Определение 2. *Статистической моделью естественного языка называется семейство функций*

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

где R^P — пространство параметров, Ω^* — пространство строк, составленных из слов словаря Ω, а $[0, 1]$ — интервал оценки вероятности появления строки в языке.

Качество модели оценивается по тестовым строкам текста величиной перплексии.

Определение 3. *Перплексией называется величина*

$$\mathrm{PP} = \frac{1}{\sqrt[k]{\Pr(w_1 w_2 \ldots w_k)}}.$$

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

3. Модель n-грамм

При отсутствии предположений о длине предыстории и о вероятности $\Pr(w_k \mid w_1^{k-1})$ число параметров будет равно числу всевозможных строк языка и будет бесконечнорастущим с ростом длины строки. В методе n-грам две предыстории считаются одинаковыми, если они оканчиваются на одинаковые $n-1$ слов.

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

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

Пример 1. *Статистическая модель биграмм задаёт следующее семейство функций:*

$$f = \Pr(w_1 w_2 \ldots w_n) = \Pr(w_n \mid w_{n-1}) \cdot \Pr(w_{n-1} \mid w_{n-2}) \cdot \ldots \cdot \Pr(w_2 \mid w_1) \cdot \Pr(w_1).$$

Число параметров модели (4) определено следующей леммой.

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

Если словарь содержит V слов, то униграммы порождают модель, имеющую $V-1$ независимых параметров: V параметров $\Pr(\tilde{w}_i)$ связаны равенством

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

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

Оценка параметров модели выполняется по коллекции текстов T. Пусть $C(w)$ — число раз, которые строка w встретилась в обучающем тексте. Тогда в случае *униграмм* максимум правдоподобия для параметра $\Pr(w)$ достигается при $\Pr(w) = \frac{C(w)}{T}$. Действительно, $V^{n-1}(V-1)$ параметров, имеющих форму $\Pr(w_n \mid w_1^{n-1})$, и $V^{n-1} - 1$ параметров более низкого порядка (по предположению индукции). Всего

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

Для случая n-грамм максимум правдоподобия равен

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

4. Модель n-грамм на классах

Пусть существует некоторая функция $\pi : \Omega \to G$, где Ω — множество слов, словарь, а G — множество классов слов. Тогда обозначим $Pr(w \mid g)$ вероятность

появления в языке слова w, если известен его класс g, а $\Pr(g_n \mid g_1^{n-1})$ — вероятность встретить слово из класса g_n после последовательности слов, имеющих форму $g_1 g_2 \ldots g_{n-1}$. Оценим только параметры вида $\Pr(g_n \mid g_1^{n-1})$ и $Pr(w \mid g)$.

Определение 5. *Модель n-грамм назовём моделью n-грамм на классах, если выполняется гипотеза*

$$\Pr(w_k \mid w_1^{k-1}) = \Pr(w_k \mid g) \Pr(g_k \mid g_1^{k-1}), \text{где } k = 1, \ldots, n. \quad (7)$$

Пример 2. *Статистическая модель биграмм на классах задаёт следующее семейство функций:*

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

Число параметров модели (7) определено следующей леммой.

Утверждение 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 $\Pr(g_n \mid g_1^{n-1})$, что доказывается аналогично Лемме 1, и $V - C$ параметров вида $Pr(w \mid g)$, так как всего таких вероятностей V ($Pr(w_i \mid g_i), i \in \{1, \ldots, V\}$), но для каждого класса $g \in G$ выполняется равенство:

$$\sum_{w:\pi(w)=g} \Pr(w \mid g) = 1. \quad (8)$$

Опишем алгоритм построения функции π на примере биграмм. Пусть $T = (t_1, t_2, \ldots, t_T)$ — текст, причём все слова содержатся в словаре Ω. Функция правдоподобия данного текста тогда равна

$$L(T) = \Pr(T) = \prod_{x,y \in \Omega} \Pr(y \mid x)^{C(xy)}, \quad (9)$$

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

$$L(T) \to \max_{\pi}. \tag{10}$$

Утверждение 3. *Задача максимизации (10) 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:

$$\log L(T) = \sum_{x,y \in \Omega} C(xy) \cdot \log \Pr(y \mid x). \tag{11}$$

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

$$\Pr(w_i \mid w_{i-1}) = \frac{C(w_i)}{C(\pi(w_i))} \cdot \frac{C(\pi(w_{i-1})\pi(w_i))}{C(\pi(w_{i-1}))}, \tag{12}$$

где $C(w_i)$ — число раз, которые слово w_i встретилось в обучающей выборке, а $C(\pi(w))$ — число раз, которое слова из класса $\pi(w)$ встретились в выборке, аналогично $C(\pi(w_x)\pi(w_y))$ — число пар вида "$\pi(w_x)\pi(w_y)$", встретившихся в выборке. Подставим теперь это выражение в функцию правдоподобия и преобразуем:

$$\log L(T) = \sum_{x,y \in \Omega} C(xy) \cdot \log\left(\frac{C(y)}{C(\pi(y))} \cdot \frac{C(\pi(x)\pi(y))}{C(\pi(x))} \right) \tag{13}$$

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

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

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

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

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

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

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

$$F_\pi = \sum_{g,h\in G} C(gh)\cdot\log C(gh) - 2\sum_{h\in G} C(h)\cdot\log C(h). \qquad (14)$$

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

1. Для всех слов $w\in\Omega$ из словаря $G(w)=1$ инициализировать набор классов.

2. Для всех начал слов $i=1\ldots n$: и всех классов $c\in G$ повторять следующие шаги F_π не перестанет увеличиваться.

3. Переместить слово w в класс c, запомнив его предыдущий класс.

4. Вычислить изменения F_π для этого перемещения в c. Переместить слово w назад в его предыдущий класс.

5. Переместить слово w в класс, который больше всего увеличивает F_π, или никуда не перемещать, если увеличения ни на каком перемещении не происходит.

Вышеописанный алгоритм сходится к локальному максимуму F_π. Это утверждение следует из того, что на каждом шаге значение F_π увеличивается.

5. Дисконтная модель

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

$$\Pr(S) = \frac{s}{A}. \qquad (15)$$

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

$$s' = d_s \cdot s, \qquad (16)$$

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

$$\Pr_{\text{discount}}(S) = \frac{s'}{A} = \frac{d_s \cdot s}{A}. \qquad (17)$$

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 $A = \sum_{s\geq 1} c_s \cdot s$. Таким образом, мы

перераспределили оценки вероятности между событиями, вероятность всех не встретившихся в обучении слов равна $1-\frac{1}{A}\sum_{s\geq 1} d_s \cdot c_s \cdot s$. Если c_0 — число таких событий, то оценка вероятности каждого из них равна

$$\frac{1}{c_0}\left(1-\frac{1}{A}\sum_{s\geq 1} d_s \cdot c_s \cdot s\right). \tag{18}$$

Сравним вышеописанную модель с моделями, предложенными в [6, 7, 8].

6. Дисконтная модель Гуда-Тьюринга

В статье [6] предлагается следующая стратегия выбора множителя:

$$d_s = (s+1)\frac{c_{s+1}}{s \cdot c_s}. \tag{19}$$

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.

7. 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:

$$d_s = \begin{cases} \dfrac{(s+1)\dfrac{c_{s+1}}{s \cdot c_s} - (k+1)\dfrac{c_{k+1}}{c_1}}{1-(k+1)\dfrac{c_{k+1}}{c_1}}, & 1 \leq s \leq k \\ 1, & s > k \end{cases} \tag{20}$$

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

Таблица 1. Некоторые свойства обучающей и тестовой выборки.

Обучающая выборка			
Размер	Кол-во униграмм	Кол-во биграмм	Кол-во триграмм
	199260	1576969	2462694
Тестовая выборка			
Все	4901	6942	6215
Новые	489	4063	5549

8. Модель абсолютного дисконтирования

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

$$d_s = \frac{s-m}{s}. \tag{21}$$

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

$$m = \frac{c_1}{\sum_{s\geq 1} c_s}. \tag{22}$$

9. Вычислительный эксперимент

Для проведения вычислительного эксперимента был использован корпус данных ISABASE для обучения статистических моделей.

Таблица 2. Перплексия тестовой выборки.

Модель дисконтирования	n-граммы	n-граммы на классах
Без дисконтирования	∞	∞
Гуд-Тьюринг	∞	∞

Катц	9560	7420
Абсолютное	10070	8210

Таким образом, в обучающей выборке биграммами покрыто 0,00397% , а триграммами — $3 \cdot 10^{-8}\%$.

Для теста использовались данные для из того же корпуса. Можно обратить внимание, что, несмотря на большой размер обучающей выборки, число новых слов довольно велико, а биаграммы и триграммы в основном новые и есть.

Количество невидимых слов, которые мы ожидаем увидеть — внешний параметр модели. Настраивался он методом скользящего контроля. Получили четко выраженный минимум при 11500 словах. Это значение и было использовано для подсчета перплексии тестовой выборки.

Видим, что не все методы одинаково хорошо проявили себя в эксперименте. вычислительный эксперимент показал, что методы без дисконтирования не справились по очевидной причине присутствия в тестовой выборке новых слов по сравнению с обучающей, а метод с дисконтированием Гуда-Тьюринга не справился, потому что среди n-грамм был разрыв в частотах, а в процессе теста встретились слова, которые были оценены нулевой вероятностью.

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 документе, однако они могут работать неадекватно, если велика доля ненадёжно обученных параметров. Также недостатком моделей Гуда-Тьюринга и Катца является их неустойчивость.

Литература

[1] Huang X., Acero A., Hon H.. Spoken Language Processing, A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 2001.

[2] Jelinek F. Statistical Methods for Speech Recognition. Cambridge: The MIT Press, 1997.

[3] Gotoh Y., Renals S. Statistical language modelling // ELSNET Summer School, Lecture Notes in Computer Science. Berlin:Springer, 2000. Vol. 2705. Pp 78–105.

[4] Young S., Bloothooft G. *Corpus-Based Methods in Language and Speech Processing*. Dordrecht: Kluwer Academic Publishers, 1997.

[5] Brown P.F., Della Pietra V. J., deSouza P.V., Mercer R. L. Class-based n-gram models of natural language // *Proceedings of the IBM Natural Language ITL*. Paris, 1990. Pp 283–298.

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

[7] Katz S.M. *Estimation of probabilities from sparse data for the language model*

component of a speech recognizer // IEEE Transactions on Acoustics, Speech and Signal Processing, 1987. Vol 35(3). Pp. 400–401.

[8] Ney H., Essen U., Kneser R. *On structuring probabilistic dependencies in stochastic language modelling* //Computer Speech and Language, 1994. Vol. 8(1). P. 38.

E. A. Budnikov, Moscow Institute of Physics and Technology
V. V. Strijov, Computing Center of the Russian Academy of Sciences

Estimating probabilities of text strings in collections of documents

Consider the problem of estimating the probabilities of strings in a document. To solve the problem, the model of n-grams is used. The n-gram classes is proposed to solve the estimation problem the large number of model parameters. Three discount models: Good-Turing, Katz and absolute discounting are used to solve the problem of zero probability of strings. The proposed model is illustrated by computational experiments on real data.

Keywords: statistical model, discount model, n-gram class, Good-Turing model, Katz model, absolute discounting.