УДК 519.766.4

А.П. Зыков

МЕТОД СГЛАЖИВАНИЯ ВЕРОЯТНОСТЕЙ N-ГРАММ НА ОСНОВЕ МОДЕЛИРОВАНИЯ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ ИХ ВСТРЕЧАЕМОСТИ

Зыков А.П. **Метод сглаживания вероятностей n-грамм на основе моделирования математического ожидания их встречаемости.**
Аннотация. В работе предлагается метод сглаживания n-граммной модели языка, в основе которого лежит моделирование функции математического ожидания вероятности встречаемости *n*-грамм. Вместо дисконтирования максимальной вероятности n-грамм предлагается увеличение мощности обучающего множества на ожидаемое число n-грамм, отсутствующих в обучающей базе текстов. Для моделирования этого числа функция математического ожидания вероятности встречаемости экстраполируется к нулевой частоте. На основе статистического анализа текстов построена модель функции математического ожидания встречаемости.
Ключевые слова: модель языка, метод сглаживания.

Zykov A.P. **N-gram smoothing based on modeling of expectation of n-gram occurrence.**
Abstract. It is shown that expectation of n-gram frequency of occurrence depends on the size of the training set and the size of the dictionary, which has been formed on the basis of this set. A method for smoothing of *n*-gram language model regarding probabilities of n-grams of lower order is proposed. This approach is based on the modeling of expectation function of n-gram occurrence probability. We suggest enlarging the size of the training set on the expected number of unseen n-grams instead of discounting maximum n-gram probability. To model the number of unseen n-grams expectation function of n-gram frequency of occurrence is extrapolated to zero frequency. Expectation function is modeled by the statistical analysis of occurrences of words in texts.
Keywords: language model, smoothing techniques.

1. Введение. В классической структуре систем распознавания речи [1] можно выделить этап поиска наиболее вероятной фразы - лингвистический процессор. Для обеспечения работы лингвистического процессора необходимо обладать моделью языка. Для построения этой модели в системах распознавания речи используются статистические методы [2, 3, 4]. Они основаны на том, что вероятность того или иного фрагмента речи, который без ограничения общности можно назвать предложением S, как последовательности из s слов данного языка, может быть представлена в виде:

$$P(S) = \prod_s P(w_s \mid W_1^{s-1}).$$

Наиболее популярная в настоящее время модель языка - n-граммная. Построение этой модели сводится к определению вероятностей цепочек элементов языка длины n (обучение) и вычислению вероятности распознаваемого предложения S с помощью выражения:

$$P(S) = \prod_s P(w_s \mid W_{s-n+1}^{s-1}).$$

Одним из основных недостатков такой модели являются колоссальные объемы обучающих данных, которые необходимы для получения достоверных оценок вероятностей n-грамм языка. На практике создать такие объёмы данных очень трудно. Для преодоления этой трудности используются различные методы сглаживания, которые позволяют производить оценку вероятностей n-грамм в условиях недостатка или полного отсутствия данных.

В большинстве известных методов сглаживания n-грамм (абсолютное дисконтирование, Джелинека-Мерсера, Виттена-Белла, Каца, Кнезера-Нея, модифицированный аддитивный метод) для оценки вероятностей $P(w_s \mid W_{s-n+1}^{s-1})$ используют значения вероятностей n-грамм более низкого порядка [3].

В данной работе, как и в перечисленных выше, рассматривается метод, в котором для оценки вероятностей n-грамм так же используются вероятности n-грамм более низкого порядка. Но гипотезы о вероятности отсутствующих n-грамм строятся на основе статистического анализа встречаемости n-грамм в обучающих выборках различной мощности и с различной мощностью словаря, при учете, что встречаемость n-грамм – случайная величина, зависящая от этих параметров.

2. Обоснование предлагаемого метода. Пусть обучающая база содержит все возможные n-граммы данного языка с той частотой, с которой они встречаются в реальной речи, тогда вероятность появления слова w_n с условием, что перед ним имеет место последовательность слов $W_1^{n-1} = w_1 ... w_{n-1}$, можно определить следующим образом:

$$P_{ML}(w_n \mid W_1^{n-1}) = \frac{r}{N(W_1^{n-1})} \quad (1)$$

где $r = C(W_1^n)$ - количество n-грамм W_1^n в обучающем множестве; $N(W_1^{n-1}) = \sum_{w_n} r_n$ - общее количество n-грамм с предысторией W_1^{n-1}.

Здесь и далее вместо обозначения $P(w_s \mid W_{s-n+1}^{s-1})$ будем использовать обозначение $P(w_n \mid W_1^{n-1})$. Как уже отмечалось выше, проблема недостатка обучающих данных, не позволяющая получить достоверные оценки вероятностей (1) для всего множества n-грамм, решается применением методов сглаживания. Большинство известных на данный момент методов сглаживания может быть сведено к выражению, предложенному Кацем [3]:

$$P_{smooth}(w_n \mid W_1^{n-1}) = \begin{cases} P_d(w_n \mid W_1^{n-1}) & при \quad r \neq 0 \\ \alpha \cdot P_{smooth}(w_n \mid W_2^{n-1}) & при \quad r = 0 \end{cases} \qquad (2)$$

где $P_d(w_n \mid W_1^{n-1}) = d \cdot P_{ML}(w_n \mid W_1^{n-1})$ - дисконтированное значение максимального значения вероятности данной n-граммы (1); α - нормировочный коэффициент; $d = r^* / r$ – дисконтирующий множитель; $P_{smooth}(w_n \mid W_2^{n-1})$ - сглаженная вероятность n-граммы с предысторией на одно слово слева меньше, чем у исследуемой n-граммы. Цель дисконтирования максимальной вероятности n-граммы

$$P_d(w_n \mid W_1^{n-1}) = d \cdot P_{ML}(w_n \mid W_1^{n-1}) = \frac{r^*}{N} \qquad (3)$$

состоит в том, чтобы уменьшить вероятность (1) для выделения части вероятности на несуществующие в обучающем множестве n-граммы, но которые могут появиться в реальных условиях распознавания речи. В выражении (3) r^* - это дисконтированное по тем или иным соображениям значение количества встреч данной n-граммы. Основы такого подхода к сглаживанию вероятностей модели языка были изложены в работе [5] и известны как оценка Гуда-Тьюринга. Дисконтированное значение r^* оценивается следующим образом:

$$r^* = N\frac{r+1}{N+1}\frac{E_{N+1}[k_{r+1}]}{E_N[k_r]} \approx (r+1)\frac{k_{r+1}}{k_r}, \quad (4)$$

где k_r - количество n-грамм, которые в обучающем множестве встретились r раз; $E_N[k_r]$ – математическое ожидание этого числа.

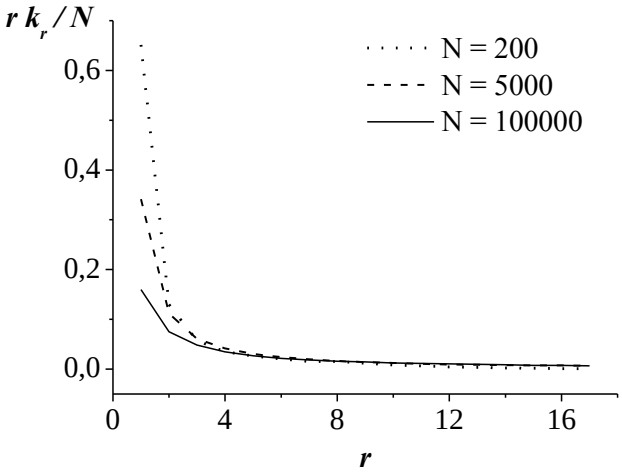

Рис. 1. Вероятность встречаемости слов для различных объёмов обучающего множества.

Однако, приближённая оценка r^* (4), которая применяется при построении модели языка со сглаживанием по методам Каца [6] и аналогичным ему, для случаев, когда $k_r \neq 0$, а $k_{r+1} = 0$, даёт значение вероятности равное нулю для n-грамм, которые реально существуют в обучающем множестве. Кроме того, такие методы плохо работают при малых значениях r, что было показано, например, в работе [7]. Причины этого следует искать в применимости приближения (4). Замена математического ожидания $E_N[k_r]$ реальным количеством n-грамм k_r, которое получено в результате обучения модели языка, является необоснованным. Во-первых, значения k_r, полученные при обучении, зависят от выбора текстов для

обучающего множества, т.е. от выборки, на которой строится модель. Во-вторых, как будет показано ниже, математические ожидания этих величин зависят от мощности обучающей выборки и от мощности словаря, который получен на этой выборке.

Рассмотрим подробнее изложенный выше подход, предложенный Гудом и Тьюрингом. Перегруппировка членов в выражении (4) позволяет получить равенство:

$$\frac{r^* \cdot E_N[k_r]}{N} = \frac{(r+1) \cdot E_{N+1}[k_{r+1}]}{N+1}.$$

Выражение $r \cdot k_r$ имеет смысл суммарного количества n-грамм, которые в обучающей базе текстов встретились по r раз каждая. Можно рассмотреть множество элементарных событий, заключающихся в том, что случайно выбранная n-грамма встретится в обучающем множестве r раз, и рассмотреть случайную величину K_r, которая равна общему количеству n-грамм, встретившихся в данном обучающем множестве r раз. Назовём эту величину встречаемостью n-грамм. Её вероятность определяется следующим образом:

$$P(K_r) = P\{K_r = r \cdot k_r\} = \frac{r \cdot k_r}{N}. \quad (5)$$

В рамках такого подхода суть дисконтирования сводится к снижению встречаемости n-грамм и, как следствие, снижению значений вероятности (5), и выделению части значений вероятностей на n-граммы с нулевой встречаемостью. Далее, дисконтирование, например, в методе Каца фактически основано на предположении о равновероятности n-грамм с встречаемостью, соответствующей нулю, и с встречаемостью, соответствующей единице. Насколько обоснован такой подход можно оценить, рассмотрев непрерывную функцию, которая при натуральных значениях r совпадала бы с вероятностью (5). На Рис. 1 приведены графики таких функций, полученные для различных мощностей обучающего множества. Судя по характеру изменения поведения кривых оценка вероятности n-грамм с нулевой встречаемостью, предложенная в методе Каца, достижима лишь при очень больших значениях мощности обучающего множества. На практике же при ограниченной обучающей базе приходится оценивать вероятности n-грамм для небольших значений N, и чем выше порядок n-грамной модели, тем меньшие значения N могут встречаться.

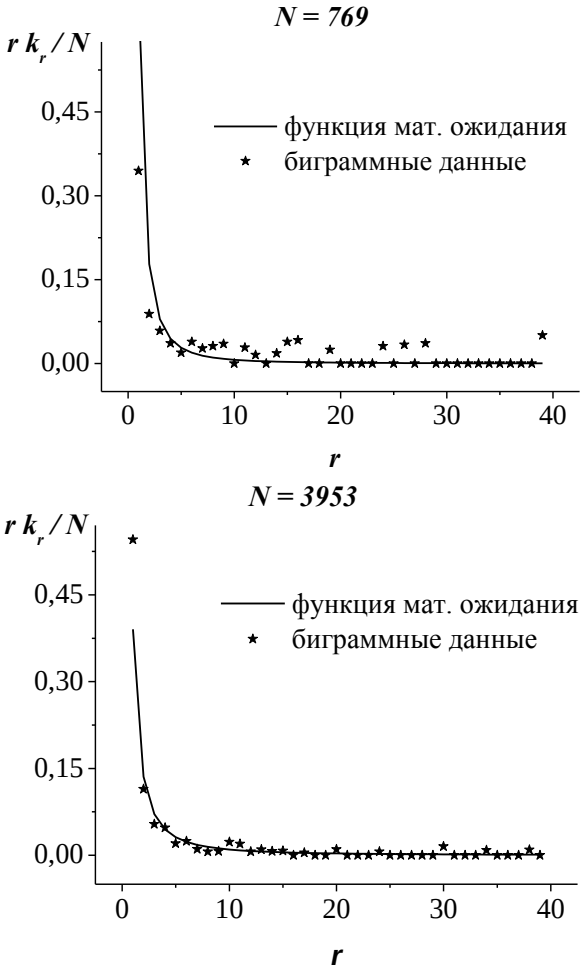

Рис. 2. Вероятности встречаемости биграмм и функции математического ожидания встречаемости униграмм для различных мощностей обучающего множества.

Предлагаемый в данной работе метод отличается от аналогичных двумя основными положениями. Во-первых, вместо дисконтирования максимальной вероятности n-грамм (1) предлагается увеличить их

общее количество на величину N_0 ожидаемого количества n-грамм, отсутствующих в обучающем множестве. Во-вторых, предлагается определять вероятности n-грамм не по их реально наблюдаемому количеству в обучающем множестве, а по математическому ожиданию этой величины:

$$r^* = [r] = \frac{N \cdot E(r)}{k_r},\ E(r) = \frac{E[K_r]}{N} = \frac{[r \cdot k_r]}{N}\ , \tag{6}$$

которое, в свою очередь, определяется по математическому ожиданию вероятности встречаемости — $E(r)$. Введём в рассмотрение непрерывную функцию $\mathcal{E}(r)$, которая при натуральных значениях r будет совпадать с дискретной функцией $E(r)$. Такой шаг позволит произвести оценку ожидаемого количества отсутствующих n-грамм. Для этого экстраполируем функцию $\mathcal{E}(r)$ до нуля, и тогда выражение $N \cdot \mathcal{E}(0)$ будет оценкой ожидаемого количества отсутствующих n-грамм.

Кроме того, в рамках предлагаемого метода сглаживания предполагается, что встречаемость n-грамм – случайная величина, не зависящая от предыстории n-грамм. Обоснованность такой гипотезы подтверждают графики на Рис. 2, где показаны вероятности встречаемости биграмм и функции математического ожидания вероятности встречаемости отдельных слов русского языка, полученные на обучающих множествах одинаковой мощности.

Предположим, что мы знаем явный вид функции, моделирующей вероятность встречаемости - $\mathcal{E}(r)$, тогда ожидаемое количество встреч некоторой n-граммы W_1^n в обучаемом множестве будет определяться следующим образом:

$$r^* = C_{ex}(W_1^n) = \begin{cases} \dfrac{N \cdot \mathcal{E}(r)}{k_r} & при \quad r \neq 0 \\ N \cdot \alpha_{ex} \cdot P_{ex}(w_n \mid W_2^{n-1}) & при \quad r = 0 \end{cases}. \tag{7}$$

Нормировочный коэффициент $\alpha_{ex}(W_1^{n-1})$ определяется из следующего равенства:

$$N \cdot \alpha_{ex} \sum_{w_n : C(W_1^n)=0} P_{ex}(w_n \mid W_2^{n-1}) =$$

$$= N \cdot \alpha_{ex} \cdot \left[1 - \sum_{w_n : C(W_1^n) \neq 0} P_{ex}(w_n \mid W_2^{n-1}) \right] = N_0 .$$

Здесь $N_0 = N \cdot \mathcal{E}(0)$ - это математическое ожидание количества n-грамм, не вошедших в обучающую базу. Тогда нормировочный коэффициент вычисляется по формуле:

$$\alpha_{ex}(W_1^{n-1}) = \frac{\mathcal{E}(0)}{1 - \sum_{w_n : C(W_1^n) \neq 0} P_{ex}(w_n \mid W_2^{n-1})}, \quad (8)$$

а сглаженная вероятность с учётом количества отсутствующих n-грамм по формуле:

$$P_{ex}(w_n \mid W_1^{n-1}) = \frac{C_{ex}(W_1^n)}{N \cdot \left[\mathcal{E}(0) + \sum_{w_n : r = C(W_1^n) \neq 0} \frac{\mathcal{E}(r)}{k_r} \right]}. \quad (9)$$

Таким образом, метод сглаживания вероятностей n-грамм будет определён, если будет построена модель функции математического ожидания случайной величины $\mathcal{E}(r)$.

3. Построение функции математического ожидания встречаемости. Как уже было показано выше (Рис. 1), функция (6) зависит от мощности обучающего множества. Для решения задачи моделирования функции (6) рассматривались обучающие последовательности слов различной мощности, полученные на нормализованных русских текстах. Мощности обучающих множеств рассчитывались по формуле:

$$N = [\exp(4.25 + k \cdot 0.25)], \quad k = 0,...30$$

Для каждого значения мощности исследовалось 1000 различных выборок. Далее выборки группировались по полученным значениям мощности словаря, и в полученных таким образом группах вычислялись математические ожидания вероятностей встречаемости n-грамм (5). Группа считалась достаточной, если в неё попадало не менее 10 выборок. В результате были получены зависимости математического ожидания встречаемости для различных значений

мощностей обучающей выборки и словаря. Пример таких зависимостей для выборок мощностью $N = 148$ представлен на рис.3.

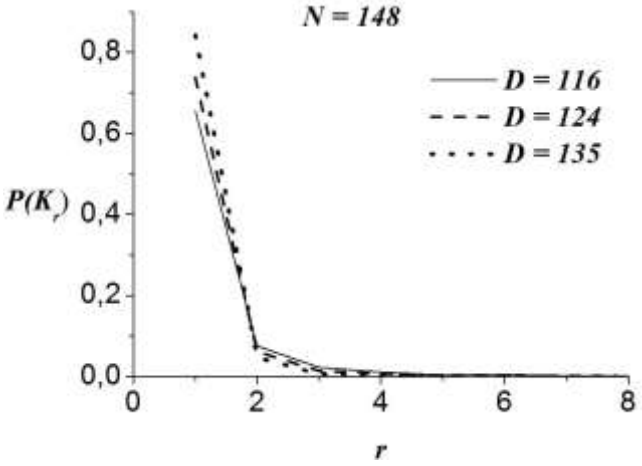

Рис. 3. Математическое ожидание вероятности встречаемости слов для различных мощностей словаря D при объёме обучающих выборок N = 148.

Основной вопрос при построении модели – выбор функции, интерполирующей полученные зависимости. Поскольку одним из необходимых условий выбора функции является возможность её экстраполяции до нуля, то интерполяция полиномами в данном случае не подходит, т.к. желательно сохранить характер функции за пределами области интерполяции. Кроме того, желательно, чтобы моделирующая функция имела как можно меньше независимых параметров, поскольку сами эти параметры будут функциями мощностей обучающего множества и словаря. В результате подбора в качестве моделирующей функции была выбрана гипербола p-й степени:

$$\mathcal{E}_{N,D}(r) = b \cdot (\sqrt[p]{1 + r^p} - r).$$

Такие же ограничения накладываются и на выбор остальных моделирующих функций. В результате были получены следующие зависимости:

$$b(N,D) = C_b + \frac{B_b}{A_b + D/N},$$

$$p(N,D) = C_p + \frac{B_p}{A_p + D/N},$$

$$B_b(N) = b_b\left(N_{ob} - N + \sqrt[d]{a_b + \left|N - N_{ob}\right|^d} \right),$$

$$C_b(N) = c_b(-N + \sqrt[f]{g_b + N^f}),$$

$$B_p(N) = B_o - t(N - N_{op})(q+1) + \sqrt{q(2t^4 + t^2 + 1)(N - N_{op})^2 + \beta(t^2+1)(t^2 - k)},$$

$$C_p(N) = b_p(-N + \sqrt[s]{a_p + N^s}).$$

Подбор параметров моделирующих функций проводился методами нелинейной регрессии на полученных статистических данных.

4. Заключение. В работе рассмотрен метод сглаживания вероятностей n-грамм, основанный на моделировании математической вероятности встречаемости n-грамм. На основе статистического анализа построена модель функции математического ожидания встречаемости n-грамм. Необходимо отметить, что эффективность того или иного метода сглаживания следует оценивать на практике, что и предполагается сделать, как продолжение данной работы. Представляет интерес провести подобные исследования для других языков и сравнить полученные результаты. Оценка количества отсутствующих в обучающем множестве n-грамм позволит оценить влияние этого параметра на качество модели сглаживания. Однако уже теоретический анализ позволяет сравнить различные методы сглаживания и оценить условия их применимости, что и было сделано для метода сглаживания Каца и предлагаемого метода. В теоретической части исследования необходимо оценить влияние предыстории n-грамм на встречаемость, оценить существенность влияния таких параметров как мощность обучающего множества и мощность словаря на эту случайную величину.

Литература

1. F. Jelinek. Continuous speech recognition by statistical methods. // Proc. IEEE, vol. 64, pp. 532-556, Apr. 1976.
2. Кипяткова И.С., Карпов А.А. Разработка и исследование статистической модели русского языка // Труды СПИИРАН. Вып. 12, СПб.: Наука, 2010, С. 35-49.
3. Chen S.F. and Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling. // Computer science group, Harvard University, Cambridge, Massachusetts, TR-8-98, August, 1998.
4. Ronald Rosenfeld. Two decades of statistical language modeling: where do we go from here? / School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
5. Good, I.J. 1953. The population frequencies of species and the estimation of population parameters. // Biometrika, 40 (3 and 4):237-264.
6. Katz, Slava M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. / IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-35 (3): 400-401, March.
7. T.Kawabata, M.Tamoto. Back-off method for N-gram smoothing based on binomial posteriori distribution. // NTT Basic Laboratories, 3-1 Morinosato-Wakamiya, Atsugi-Shi 243-01, Japan.

Зыков Александр Павлович – старший научный сотрудник отдела речевых технологий ООО «Стэл КС». Область научных интересов: автоматическое распознавание речи, построение моделей языка, численное моделирование гидродинамики реагирующих потоков. Число научных публикаций – 14. zykov_ap@stel.ru; 105082, Москва, ул. Б.Почтовая 55/59; р.тел./факс +7(495)775-51-23

Zykov Aleksandr Pavlovich - senior researcher, Department of speech technologies, "Stel CS" Ltd. Research interests: automatic speech processing, language modeling, numerical modeling of reacting flow. The number of publications — 14. zykov_ap@stel.ru; "Stel CS" Ltd, 55/59 B. Pochtovaja st., Moscow, 105082, Russian Federation, office phone/fax +7(495)775-51-23.

Поддержка исследований. Данное исследование выполняется в рамках научно исследовательских работ проводимых в компании «Стэл Компьютерные Системы».

Рекомендовано отделом речевых технологий, начальник отдела Леднов Д.А, канд. техн. наук.
Статья поступила в редакцию 05.07.2011.

РЕФЕРАТ

Зыков А.П. **Метод сглаживания вероятностей n-грамм на основе моделирования математического ожидания их встречаемости.**

В работе рассматривается статистический подход к построению метода сглаживания моделей языка. Для этого рассматривается случайная величина - встречаемость n-грамм в обучаемом множестве. Показано, что математическое ожидание встречаемости n-грамм зависит от мощности обучающего множества и размера словаря, полученного на этом множестве. Предлагается метод сглаживания n-граммной модели языка с учётом вероятностей n-грамм более низкого порядка. В основе предлагаемого метода лежит моделирование функции математического ожидания вероятности встречаемости n-грамм. Предлагается определять вероятности n-грамм не по их реально наблюдаемой частоте в обучающем множестве, а по математическому ожиданию этой величины. Вместо дисконтирования максимальной вероятности n-грамм предлагается увеличивать их общее количество на величину ожидаемого количества n-грамм, отсутствующих в обучающем множестве. Для моделирования этого числа функция математического ожидания вероятности встречаемости n-грамм экстраполируется к нулевой частоте. При построении метода сглаживания предполагается, что встречаемость n-грамм – случайная величина, не зависящая от предыстории n-грамм. На основе статистического анализа слов в базе русских текстов построена модель функции математического ожидания встречаемости слов.

SUMMARY

Zykov A.P. **N-gram smoothing based on modeling of expectation of n-gram occurrence.**

In this paper statistical approach to language model smoothing is considered. For this purpose n-gram occurrence is used as random variable. It is shown that expectation of n-gram frequency of occurrence depends on the size of the training set and the size of the dictionary, which has been formed on the basis of this set. A method for smoothing of n-gram language model regarding probabilities of n-grams of lower order is proposed. This approach is based on the modeling of expectation function of n-gram occurrence probability. We propose estimating n-gram probabilities by frequency expectation and not by its actual frequency in the training set. We suggest enlarging the size of the training set on the expected number of unseen n-grams instead of discounting maximum n-gram probability. To model the number of unseen n-grams expectation function of n-gram frequency of occurrence is extrapolated to zero frequency. This approach to n-gram smoothing presupposes that n-gram frequency is a random variable not depending on n-gram history. Expectation function is modeled by the statistical analysis of occurrences of words in Russian texts.