1977

Кибернетика

ОТДЕЛЬНЫЙ ОТТИСК

В. В. НЕШИТОЙ

РАСПРЕДЕЛЕНИЕ КЛЮЧЕВЫХ СЛОВ В ТЕКСТЕ

УДК 519.2:801

При решении большого количества задач, связанных со статистической обработкой текста, необходимо знать закон распределения разных слов по частоте их употребления в тексте, а также аналитическую зависимость, описывающую кривую роста новых слов в тексте. При этом новым считается любое из разных слов при первом его появлении от начала текста.

Обозначим через Y дискретную случайную величину — число новых (разных) событий, которые наступают в результате x независимых испытаний. При этом $x = 0, 1, 2, \ldots$; $Y = 0, 1, 2, \ldots, n$, где n — число несовместных событий, составляющих полную группу (n может быть как конечным, так и бесконечным). Если вероятности каждого из n событий заданы, т. е. задан закон распределения вероятностей разных событий, то теоретически может быть найдена функциональная зависимость между числом произведенных испытаний x и математическим ожиданием (средним значением) числа наступивших новых событий $M[Y]$. Другими словами, может быть найдено математическое ожидание $M[Y(x)]$ случайной функции $Y(x)$, т. е. кривая роста новых событий. Примером такой кривой является функциональная зависимость между числом словоупотреблений в некоторой выборке (длиной текста) и средним значением числа разных слов (объемом словаря). Кривая роста новых слов может быть рассчитана теоретически, если известен закон распределения разных слов в тексте.

На практике может возникнуть необходимость решить обратную задачу — по заданной кривой роста новых событий установить закон распределения разных событий.

В настоящей работе ставятся следующие задачи:

— исследовать возможности использования кривых роста новых событий для установления законов распределения разных событий;

— найти аналитическую зависимость между объемом выборки и средним значением объема словаря;

— установить закон распределения ключевых слов в тексте.

ВЕРОЯТНОСТЬ ПОЯВЛЕНИЯ НОВОГО СОБЫТИЯ И МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ СЛУЧАЙНОЙ ФУНКЦИИ

Рассмотрим полную группу попарно несовместных событий $A_1, A_2, \ldots, A_k, \ldots, A_n$, вероятности которых заданы и соответственно равны $p_1, p_2, \ldots, p_k, \ldots, p_n$. Пусть производится $i = x$ независимых испытаний, в каждом из которых может появиться одно из n разных событий, например событие A_k. Найдем вероятность того, что событие A_k появится первый раз в $(i+1)$-ом испытании и, следовательно, не появится при первых i испытаниях $(i = 0, 1, 2, \ldots, x)$. Другими словами, найдем вероятность появления нового события A_k в $(i+1)$-ом испытании. Она определится по формуле

$$P_{i=x}(A_k) = p_k (1 - p_k)^i, \qquad (1)$$

где i — число произведенных испытаний.

Рассмотрим далее новое событие A, представляющее собой сумму n новых событий $(A = A_1 + A_2 + \ldots + A_k + \ldots + A_n)$, и найдем вероятность появления этого нового события в $(i+1)$-ом испытании. Для несовместных событий эта вероятность равна

$$P_{i=x}(A) = \sum_{k=1}^{n} P_{i=x}(A_k) = \sum_{k=1}^{n} p_k (1 - p_k)^i. \qquad (2)$$

Учитывая далее, что математическое ожидание числа новых событий, наступающих при одном $(i+1)$-ом испытании, равно вероятности появления нового события в этом испытании, т. е. $M[Z_{i+1}] = P_i(A)$, найдем математическое ожидание числа новых событий, наступающих при $i = x$ испытаниях:

$$M[Y(x)] = \sum_{i=0}^{x-1} M[Z_{i+1}] = \sum_{i=0}^{x-1} P_i(A),$$

или с учетом (2)

$$M[Y(x)] = \sum_{i=0}^{x-1} \sum_{k=1}^{n} p_k (1 - p_k)^i =$$

$$\sum_{k=1}^{n} \sum_{i=0}^{x-1} p_k (1 - p_k) = \sum_{k=1}^{n} [1 - (1 - p_k)^x]. \qquad (3)$$

Выражение в квадратных скобках в (3) обозначает вероятность того, что k-е событие при x испытаниях произойдет хотя бы один раз.

Введем в рассмотрение непрерывную плотность распределения вероятностей $p(t)$, аппроксимирующую вероятности p_k. Она должна удовлетворять условию

$$p_k = \int_{k-1}^{k} p(t)\,dt = p(t); \quad k-1 \leqslant t < k. \tag{4}$$

На основании условия (4) формулы (2) и (3) можно выразить приближенно интегралами:

$$P_{i=x}(A) = \int_0^n p(t)\,[1-p(t)]^x dt, \tag{5}$$

$$M[Y(x)] = \int_0^n \{1-[1-p(t)]^x\}\,dt. \tag{6}$$

Если число разных событий n велико и вероятности отдельных событий малы, то при этих условиях будут малы также значения плотности распределения $p(t)$. В этом случае имеем приближенное равенство $[1-p(t)]^x \approx e^{-xp(t)}$, с учетом которого формулы (5) и (6) перепишутся в виде

$$P_x(A) = \int_0^n \frac{p(t)}{e^{xp(t)}}\,dt, \tag{5'}$$

$$y = \int_0^n \left(1 - \frac{1}{e^{xp(t)}}\right) dt. \tag{6'}$$

Здесь x и y — величины непрерывные. Таким образом, получена непрерывная функция $y = f(x)$ (6′), график которой близок к ломаной $M[Y(x)]$ (3). Назовем эту непрерывную функцию также математическим ожиданием случайной функции $Y(x)$. Как было показано в работе [1], замена ломаной (3) на приближенную плавную кривую (6′) дает возможность вычислять величину $M[Y]$ с ошибкой, не превышающей 0,5 события.

Из формул (3) и (6′) следует, что для нахождения математического ожидания случайной функции необходимо и достаточно знать вероятности всех событий, составляющих полную группу, или плотность распределения $p(t)$. При этом разные события могут быть упорядочены по любому правилу. Но эти формулы не позволяют решать обратную задачу, а именно: по известному математическому ожиданию случайной функции находить закон распределения вероятностей событий, составляющих полную группу. Примером такой задачи является установление закона распределения разных слов в тексте по известной аналитической зависимости между длиной текста x и объемом словаря $y = f(x)$.

Между вероятностью появления нового события и математическим ожиданием случайной функции существует определенная связь. Дифференцируя выражение (6′) по x, получим

$$\frac{dy}{dx} = \frac{d}{dx}\int_0^n \left(1 - \frac{1}{e^{xp(t)}}\right) dt =$$

$$= \int_0^n \frac{d}{dx}\left(1 - \frac{1}{e^{xp(t)}}\right) dt = \int_0^n \frac{p(t)}{e^{xp(t)}}\,dt,$$

или

$$\frac{dy}{dx} = P_x(A). \tag{7}$$

Таким образом, в случае большого числа несовместных событий, составляющих полную группу, первая производная от математического ожидания $y = f(x)$ случайной функции $Y(x)$ при данном значении x равна вероятности появления нового события $P_x(A)$.

РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ НОВЫХ СОБЫТИЙ

Пусть производятся независимые испытания, в каждом из которых появляется какое-нибудь из n несовместных событий, составляющих полную группу. Любое из этих событий при первом его появлении от начала испытаний будем считать новым событием. Если произвести несколько серий испытаний, то порядок появления новых событий в каждой серии будет разный. Это значит, что k-ым по порядку новым событием в каждой серии может наступить любое событие полной группы. Обозначим через $\bar{p}_1$ среднее значение (математическое ожидание) вероятностей новых событий, наступающих во всех сериях первыми от начала испытаний; через $\bar{p}_2$ — среднее значение вероятностей новых событий, которые наступают вторыми от начала испытаний; через $\bar{p}_k$ — среднее значение вероятностей новых событий, которые наступают k-ми от начала испытаний.

Так как в ходе испытаний события, составляющие полную группу, должны наступать в среднем в порядке убывания их вероятностей

то для средних вероятностей новых событий должно выполняться условие

$$\bar{p}_1 \geqslant \bar{p}_2 \geqslant \ldots \geqslant \bar{p}_k \geqslant \ldots \geqslant \bar{p}_n. \qquad (8)$$

Введем непрерывную плотность распределения $\bar{p}(y)$, удовлетворяющую условиям

$$\bar{p}_k = \int_{k-1}^{k} \bar{p}(y)\,dy = \bar{p}(y),\ k-1 \leqslant y < k;\ \frac{d\bar{p}(y)}{dy} \leqslant 0. \qquad (9)$$

Здесь $\bar{p}(y)$ есть невозрастающая средняя плотность распределения вероятностей новых событий.

Пусть в результате x произведенных испытаний наступает в среднем y разных событий. Тогда вероятность появления нового события A в точке x непрерывной кривой $y = f(x)$ согласно (7) будет равна $P_x(A) = \frac{dy}{dx}$, а накопленная вероятность y наступивших событий на основании (8) и (9) (события наступают в порядке убывания их средних вероятностей!) определится по формуле

$$P_x(\bar{A}) = \bar{p}_1 + \bar{p}_2 + \ldots + \bar{p}_{k=y} = \int_0^y \bar{p}(t)\,dt,$$

где $\bar{A}$ — событие, противоположное событию A.

Для событий, составляющих полную группу, справедливо равенство $P_x(\bar{A}) = 1 - P_x(A)$ или с учетом двух предыдущих формул

$$\int_0^y \bar{p}(t)\,dt = 1 - \frac{dy}{dx}.$$

Обозначив в последнем выражении $\int_0^y \bar{p}(t)\,dt = F(y)$, придем к дифференциальному уравнению

$$F(y) = 1 - \frac{dy}{dx}, \qquad (10)$$

где $F(y)$ есть функция распределения вероятностей новых событий, наступающих в результате x произведенных испытаний.

Из (10) дифференцированием по y получим выражение для средней плотности распределения вероятностей новых событий:

$$\bar{p}(y) = -\frac{d}{dy}\left(\frac{dy}{dx}\right) = -\frac{y''_x}{y'_x}. \qquad (11)$$

Таким образом, средняя плотность распределения $\bar{p}(y)$ связана простыми соотношениями с функцией $y = f(x)$.

Полученные выражения (10) и (11) дают возможность по известному математическому ожиданию $y = f(x)$ случайной функции $Y(x)$ находить закон распределения вероятностей новых событий. При этом кривая $y = f(x)$ должна быть неубывающей функцией, удовлетворяющей начальным условиям:

$$\begin{aligned}
&\text{а)}\ \frac{dy}{dx} = 1 && \text{при} && x = 0,\ y = 0;\\
&\text{б)}\ \frac{dy}{dx} = 0 && \text{при} && x = \infty,\ y = n;\\
&\text{в)}\ \frac{d^2y}{dx^2} < 0 && \text{при} && x > 0,\ 0 < y < n;\\
&\text{г)}\ \frac{d^2}{dy^2}\left(\frac{dy}{dx}\right) \geqslant 0 && \text{при} && x > 0,\ 0 < y < n.
\end{aligned} \qquad (12)$$

Эти условия вытекают из зависимостей (9)—(11) и являются следствием свойств закона распределения вероятностей новых событий.

Условия (12) можно интерпретировать следующим образом:

а) вероятность появления нового события в первом испытании равна единице;

б) вероятность появления нового события равна нулю при $x = \infty$, $y = n$;

в) кривая роста новых событий выпукла кверху и не имеет точек перегиба; прямая $y = n$ является горизонтальной асимптотой данной кривой;

г) средняя плотность распределения вероятностей новых событий является невозрастающей функцией.

Формулы (10) и (11) позволяют также решать обратную задачу: по заданному закону распределения вероятностей новых событий находить зависимость $y = f(x)$. Решая дифференциальное уравнение (10) относительно x, получим

$$x = \int \frac{dy}{1 - F(y)} + C. \qquad (13)$$

Постоянную интегрирования C находим из начальных условий $x = 0$ при $y = 0$. Разрешая затем полученное уравнение относительно y (если это возможно), найдем искомую функцию $y = f(x)$.

УСТАНОВЛЕНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ РАЗНЫХ СОБЫТИЙ ПО ЗАДАННОМУ МАТЕМАТИЧЕСКОМУ ОЖИДАНИЮ СЛУЧАЙНОЙ ФУНКЦИИ

Если все события, составляющие полную группу, расположить в ряд по убыванию их вероятностей, то зависимость между порядковым номером события в этом ряду и его вероятностью приближенно можно описать некоторой непрерывной невозрастающей плотностью распределения. Обозначим ее через $p(z)$, где z — порядковый номер события в ряду по убывающим вероятностям.

Пусть $F(z) = \int_0^z p(z)\,dz$ — накопленная вероятность первых z наиболее частых событий из n разных событий, составляющих полную группу, а $F(y)$ — накопленная вероятность y разных событий, наступающих при x испытаниях. Так как среди y событий могут быть как наиболее частые, так и весьма редкие события, то при $z = y$ справедливо неравенство

$$F(z) \geqslant F(y), \qquad (14)$$

причем в случае неравновозможных событий знак равенства достигается лишь при $z = y = 0$ и $z = y = n$. В случае равновозможных событий $F(z) = F(y)$ и, следовательно, $p(z) = \bar{p}(y)$.

Если в формулу (13) вместо функции распределения $F(y)$ подставить функцию распределения вероятностей событий, упорядоченных по убыванию их вероятностей $F(z)$, то получим некоторую кривую $z = \varphi(x)$, при этом на основании неравенства (14) имеем $\varphi(x) \leqslant f(x)$. Здесь величина z обозначает номер события в ряду по убывающим вероятностям и одновременно число наиболее частых событий, наступающих в результате x испытаний (т. е. принимается, что разные события наступают строго по порядку убывания их вероятностей). Функция $z = \varphi(x)$ удовлетворяет тем же начальным условиям (12), что и функция $y = f(x)$:

а) $\frac{dz}{dx} = 1$ при $x = 0,\ z = 0;$

б) $\frac{dz}{dx} = 0$ при $x = \infty,\ z = n;$

в) $\frac{d^2z}{dx^2} < 0$ при $x > 0,\ 0 < z < n;$

г) $\frac{d^2}{dz^2}\left(\frac{dz}{dx}\right) \geqslant 0$ при $x > 0,\ 0 < z < n.$

Поскольку при $x \to \infty$ $z \to n$, $y \to n$, то

$$\lim_{x\to\infty} \frac{f(x)}{\varphi(x)} = \frac{n}{n} = 1, \qquad (15)$$

т. е. функция $\varphi(x)$ является асимптотическим приближением функции $f(x)$, что записывается следующим образом: $f(x) \sim \varphi(x)$ при $x \to \infty$. Последнее равенство свидетельствует также о некоторой близости между плотностями распределения $\bar{p}(y)$ и $p(z)$. При замене $f(x)$ ее асимптотическим приближением относительная погрешность $\frac{f(x) - \varphi(x)}{f(x)}$ с увеличением x вначале быстро растет до своего максимального значения (как показывают расчеты, обычно не превышающего 30%), затем медленно убывает, стремясь к нулю при $x \to \infty$.

Таким образом, если нам известна функция $y = f(x)$, то на основании равенства (15) и начальных условий (12) можно считать, что $\varphi(x) \approx f(x)$. Из зависимости $z = \varphi(x)$ на основании формулы (10) получим выражение для функции распределения вероятностей событий, упорядоченных по убыванию их вероятностей:

$$F(z) = 1 - \frac{dz}{dx}. \qquad (16)$$

Из (16) дифференцированием по z найдем выражение для плотности распределения:

$$p(z) = -\frac{d}{dz}\left(\frac{dz}{dx}\right) = -\frac{z''_x}{z'^3_x}. \qquad (17)$$

Получаемые таким образом формулы для $F(z)$ и $p(z)$ требуют еще дальнейшей проверки по опытным данным и последующего уточнения, так как в общем случае $\bar{p}(y) \neq p(z)$ вследствие того, что $f(x) \neq \varphi(x)$. Заметим, что в случае равновозможных событий $f(x) = \varphi(x)$.

Таким образом, формулы (16) и (17) позволяют установить вид распределения вероятностей событий, если задана аналитическая зависимость $y = f(x)$.

Теперь сформулируем общее правило, которое можно применить при нахождении закона распределения разных слов в тексте [1].

1. Для нахождения закона распределения слов, заданного плотностью распределения $p(z)$, вначале необходимо найти хотя бы приближенную функциональную зависимость между длиной текста x и объемом словаря $y = f(x)$. Она должна удовлетворять начальным условиям (12).

2. На основании равенства (15) и начальных условий (12) считать, что $\varphi(x) \approx f(x)$.

3. По известной функции $z = \varphi(x)$ по формулам (16) и (17) найти функцию распределения $F(z)$ и плотность распределения вероятностей разных слов в тексте $p(z)$.

4. Произвести опытную проверку полученных формул. Для последующего уточнения найденных законов распределения необходимо использовать их свойства и возможности, заложенные в структуре формул.

5. В случае неудачного исхода выбрать более подходящую функцию $y = f(x)$ и повторить все сначала.

Если найденный закон распределения разных слов в тексте согласуется с опытом, то по формуле (6′) можно получить более точную зависимость между длиной текста и объемом словаря.

6. Закон распределения новых слов в тексте при заданной зависимости $y = f(x)$ можно найти по формулам (10) и (11). Если исходная функция $y = f(x)$ согласуется с опытом, то полученный закон распределения новых слов в тексте не требует дальнейшей проверки.

УСТАНОВЛЕНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ КЛЮЧЕВЫХ СЛОВ

Для описания в первом приближении зависимости между длиной текста x и объемом словаря $y = f(x)$ предлагается следующее дифференциальное уравнение:

$$\frac{dY}{dX} = \frac{Y}{X}(1 - u\alpha \ln Y)^{1/u}, \qquad (18)$$

где $Y = y + 1$, $X = x + 1$, $\alpha > 0$, u — некоторый параметр.

Уравнение (18) при $u \leqslant 1$ удовлетворяет всем пунктам условий (12). Оно может описывать как конечное (при $0 < u < 1$), так и бесконечное (при $u \to 0$ и $u < 0$) множество разных слов.

На основании равенства (15) и начальных условий (12) будем считать, что кривая $z = \varphi(x)$ описывается таким же дифференциальным уравнением:

$$\frac{dZ}{dX} = \frac{Z}{X}(1 - u\alpha \ln Z)^{1/u}, \qquad (18')$$

где $Z = z + 1$, z — порядковый номер слова в списке по убывающим вероятностям.

Пусть параметр $u \to 0$. Тогда уравнение (18′) преобразуется к виду

$$\frac{dZ}{dX} = \frac{Z^{1-\alpha}}{X}. \qquad (19)$$

Разделяя переменные и решая последнее уравнение, получим

$$\ln X = \frac{Z^{\alpha}}{\alpha} + C. \qquad (20)$$

Используя начальные условия $X = 1$ при $Z = 1$ ($x = 0$ при $z = 0$) найдем $C = -\frac{1}{\alpha}$. С учетом последнего равенства из (20) получим

$$X = e^{\frac{Z^{\alpha}-1}{\alpha}}. \qquad (21)$$

Заменим в (19) переменную X ее значением из (21) и запишем выражение для функции распределения вероятностей разных слов, упорядоченных по убыванию их вероятностей. Согласно (16) имеем

$$F(Z) = 1 - \frac{Z^{1-\alpha}}{e^{\frac{Z^{\alpha}-1}{\alpha}}}. \qquad (22)$$

В формулу (22) параметр α входит трижды. Это дает возможность *ввести три разных* параметра. Тогда выражение (22) можно переписать в виде

$$F(Z) = 1 - \frac{1}{Z^{d} e^{a(Z^{b}-1)}}. \qquad (22')$$

Дифференцируя (22′) по Z, найдем выражение для плотности распределения:

$$p(Z) = \frac{d + abZ^{b}}{Z^{1+d} e^{a(Z^{b}-1)}}. \qquad (23)$$

Если в (23) положить $a = 0$ (или $b = 0$), то получим закон Ципфа—Мандельброта:

$$p(z) = \frac{d}{Z^{1+d}} = \frac{d}{(z+1)^{1+d}}.$$

При больших значениях z распределение, выраженное формулами (22′) и (23), совпадает с распределением Вейбулла с параметрами a и b. Это можно показать следующим образом.

Преобразуем функцию распределения (22′) к виду

$$\ln \ln \frac{1}{1 - F(Z)} = \ln [d \ln Z + a(e^{b \ln Z} - 1)]$$

и найдем асимптоту кривой $\ln \ln \frac{1}{1 - F(Z)} = \Psi(\ln Z)$.

Тангенс угла Θ асимптоты с горизонтальной осью определится по формуле

$$\operatorname{tg}\Theta = \lim_{\ln Z \to \infty} \frac{\Psi(\ln Z)}{\ln Z},$$

или в наших обозначениях

$$\operatorname{tg}\Theta = \lim_{\ln Z \to \infty} \frac{\ln [d \ln Z + a(e^{b \ln Z} - 1)]}{\ln Z}.$$

Раскрывая неопределенность $\frac{\infty}{\infty}$ по правилу Лопиталя, получим $\operatorname{tg}\Theta = b$.

Зная $\operatorname{tg}\Theta$, найдем начальную ординату асимптоты (обозначим ее через m):

$$m = \lim_{\ln Z \to \infty} [\Psi(\ln Z) - b \ln Z],$$

или

$$m = \lim_{\ln Z \to \infty} \{\ln [d \ln Z + a(e^{b \ln Z} - 1)] - b \ln Z\} =$$

$$= \lim_{\ln Z \to \infty} \ln \frac{d \ln Z + a(e^{b \ln Z} - 1)}{e^{b \ln Z}} =$$

$$= \ln \lim_{\ln Z \to \infty} \frac{d \ln Z + a(e^{b \ln Z} - 1)}{e^{b \ln Z}} = \ln a.$$

Учитывая, что при больших значениях $Z = z + 1$ $\ln Z \approx \ln z$, запишем уравнение асимптоты в виде

$$\ln \ln \frac{1}{1 - F(z)} = \ln a + b \ln z, \qquad (24)$$

откуда

$$F(z) = 1 - \frac{1}{e^{az^b}}, \qquad (25)$$

$$p(z) = \frac{ab}{z^{1-b} e^{az^b}}, \qquad (26)$$

т. е. получили закон Вейбулла.

Этот закон впервые применил Г. Г. Белоногов для описания распределения слов в русской письменной речи [2].

Таким образом, если распределение разных слов в тексте подчиняется закону Вейбулла, то график зависимости

$$\ln \ln \frac{1}{1 - F^*(z)} = \Psi(\ln z),$$

построенный по опытной функции распределения $F^*(z)$, должен иметь вид прямой с начальной ординатой $\ln a$ и угловым коэффициентом b. Опытная проверка на литературных текстах показала, что данный график близок к прямой лишь при $z > 100 \div 500$. Чтобы учесть особенность распределения наиболее частых слов, введем третий параметр c и запишем функцию распределения (25) в виде [1]:

$$F(z) = 1 - \frac{1}{e^{a[(z+1)^b - e^{-cz}]}}. \qquad (27)$$

Из (27) дифференцированием по z найдем выражение для плотности распределения:

$$p(z) = \frac{\frac{ab}{(z+1)^{1-b}} + \frac{ac}{e^{cz}}}{e^{a[(z+1)^b - e^{-cz}]}}. \qquad (28)$$

При достаточно больших значениях z данное распределение совпадает с распределением Вейбулла.

Разрешим уравнение (27) относительно c:

$$c = -\frac{1}{z} \ln \left[(z+1)^b + \frac{1}{a} \ln (1 - F(z)) \right]. \qquad (29)$$

При вычислении значения c необходимо принимать $z = 1 \div 30$. Практически $c = 0{,}01 \div 0{,}1$.

Поскольку наиболее частыми словами, как правило, являются служебные слова, то параметр c в формулах (27) и (28) служит для более точного описания главным образом этой группы слов. Если из текста убрать все служебные слова, то распределение оставшихся разных слов должно подчиняться закону Вейбулла. Примером текста, не содержащего служебных слов, является текст, составленный из ключевых слов.

Опытная проверка показала, что для ключевых слов закон Вейбулла выполняется уже при $z > 10 \div 50$. Этим законом описывается также распределение периодических изданий по числу помещенных в них статей по данному предмету и распределение научных сотрудников по продуктивности.

ЗАВИСИМОСТЬ МЕЖДУ ДЛИНОЙ ТЕКСТА И ОБЪЕМОМ СЛОВАРЯ

При известном законе распределения разных слов в тексте, заданном плотностью распределения $p(z)$, можно более точно вычислить ожидаемый объем словаря y при заданной длине текста x. Расчет производится по формуле (6′), которая в данном случае перепишется в виде

$$y = \int_0^\infty \left(1 - \frac{1}{e^{xp(z)}}\right) dz. \qquad (30)$$

Если в формулу (30) вместо плотности распределения $p(z)$ подставить ее значение из (26), то получим интеграл, который не выражается через элементарные функции и поэтому значения y при заданных x можно рассчитать лишь путем численного интегрирования.

Предпринимались также попытки получить конкретную зависимость $y = f(x)$ на основе более простого распределения Ципфа [3], [4], но в результате не было получено приемлемых расчетных формул.

В связи с вышесказанным нам представляется целесообразным для описания кривых роста новых слов использовать простые приближенные формулы, точность которых может быть оценена с помощью уравнения (30) и плотности распределения Вейбулла (26).

Приближенные формулы можно получить путем решения дифференциального уравнения (18).

Пусть по-прежнему параметр $u \to 0$. Тогда уравнение (18) преобразуется к виду

$$\frac{dY}{dX} = \frac{Y^{1-\alpha}}{X}.$$

Решая это уравнение и используя начальные условия ($X = 1$ при $Y = 1$), найдем

$$X = e^{\frac{Y^{\alpha}-1}{\alpha}},$$

откуда

$$Y = (1 + \alpha \ln X)^{1/\alpha}.$$

Полученная формула оказалась неудобной для практических расчетов, так как ее структура не позволяет в явном виде выразить величину α через Y и X. Из общего уравнения (18) при различных значениях параметра u можно получить ряд формул для описания кривых роста новых ключевых слов в тексте. Расчеты показывают, что наиболее подходящей для этих целей является формула, полученная при $u = -1$. В этом случае выражение (18) перепишется в виде

$$\frac{dY}{dX} = \frac{Y}{X(1 + \alpha \ln Y)}.$$

Решая это дифференциальное уравнение, найдем

$$X = Y^{1+\frac{\alpha}{2}\ln Y}, \qquad (31)$$

откуда

$$Y = e^{\frac{1}{\alpha}(\sqrt{1+2\alpha \ln X}-1)}. \qquad (32)$$

Из формулы (31) можно выразить величину α через Y и X:

$$\alpha = \frac{2}{\ln Y}\left(\frac{\ln X}{\ln Y} - 1\right). \qquad (33)$$

Параметр α, вычисляемый по формуле (33) на основе опытных значений y и x, не является постоянной величиной, однако он изменяется закономерно. Если построить график зависимости α от $\ln X$, то получим прямую, уравнение которой имеет вид

$$\alpha = \alpha_0 + k \ln X, \qquad (34)$$

или

$$\alpha = \alpha_0 + k' \lg X, \qquad (34')$$

где α_0 — начальная ордината; k, $k' \approx 2{,}3\,k$ — угловые коэффициенты данной прямой.

Таким образом, если в формуле (32) считать α величиной переменной ($\alpha = \alpha_0 + k \ln X$), то она становится практически точной и содержит два параметра: α_0 и k. Эта формула справедлива при $10^3 < X < 10^8$ словоупотреблений, при этом ее погрешность на границах указанного участка не превышает $3 \div 5\%$ по сравнению с теми данными, которые можно получить по формуле (30) на основе плотности распределения Вейбулла (26).

Параметры α_0 и k являются характеристикой лексического богатства текста. Опытная проверка показывает, что тексты, относящиеся к одному жанру, характеризуются близкими значениями параметра k. При этом чем богаче данный текст в лексическом отношении, тем ниже располагается на графике прямая (34), т. е. для такого текста меньше начальная ордината α_0. Формула (32) позволяет вычислять накопленную вероятность y разных ключевых слов, образующих текст длиной x словоупотреблений, т. е. полноту словаря:

$$F(y) = 1 - \frac{dy}{dx},$$

а также находить вероятность появления нового ключевого слова в точке x кривой $y = f(x)$:

$$P_x(A) = \frac{dy}{dx}.$$

Дифференцируя (32) по X с учетом (34), после преобразований будем иметь

$$\frac{dY}{dX} = \frac{Y}{X} \cdot \frac{1 - \frac{k}{2}\ln^2 Y}{2\frac{\ln X}{\ln Y} - 1}. \qquad (35)$$

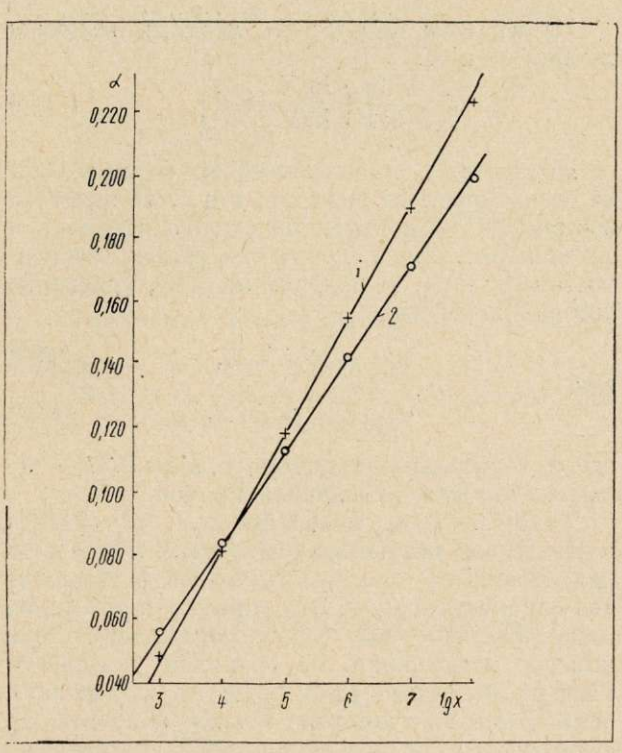

Рис. 1.

Из (35) следует, что первая производная обращается в нуль при $Y = Y_{max} = e^{\sqrt{\frac{2}{k}}}$. Из последнего выражения видно, что Y_{max} тем больше, чем меньше k. При $k = 0$ $Y_{max} = \infty$.

Из формулы (32) с учетом (34) можно найти зависимость длины текста X от объема словаря Y:

$$X = Y^{\frac{2+\alpha_0 \ln Y}{2-k\ln^2 Y}}. \qquad (36)$$

Из (36) следует, что величина Y достигает максимального значения при $X = \infty$.

В формулах (31) — (36) величина $y = Y - 1$ обозначает число разных ключевых слов, а величина $x = X - 1$ — длину текста, составленного из y ключевых слов. Если требуется найти зависимость между длиной полного текста (обозначим ее через L) и числом разных ключевых слов, то следует иметь в виду, что в этом случае $y = f(pL)$, где p — вероятность того, что взятое наугад слово текста является ключевым. При этом $pL = x$, откуда $p = x/L$. Скорость прироста новых ключевых слов в зависимости от длины текста L определится по формуле

$$\frac{dY}{dL} = \frac{Y}{L} \cdot \frac{1 - \frac{k}{2}\ln^2 Y}{2\frac{\ln pL}{\ln Y} - 1}.$$

Сравнивая последнее равенство с (35), находим

$$\frac{dY}{dL} = p\frac{dY}{dX}.$$

Чтобы показать справедливость формулы (32), на основе распределения Вейбулла с параметрами $a = 0{,}05$, $b = 0{,}60$ и $a = 0{,}10$, $b = 0{,}50$ по формуле (30) были рассчитаны значения y при заданных x. Затем по формуле (33) были вычислены значения величины α и построены графики (см. рис. 1) зависимости α от $\lg x$. Из графиков видно, что в обоих случаях почти все точки лежат на теоретической прямой $\alpha = \alpha_0 + k' \lg x$. При этом параметры прямой 1 $\alpha_0 = -0{,}0634$, $k' = 0{,}0365$. Для прямой 2 $\alpha_0 = -0{,}0335$, $k' = 0{,}0294$.

Опытным путем была установлена зависимость между параметрами α_0, k' и параметрами распределения Вейбулла a, b:

$$\alpha_0 = -0{,}0163e^{2{,}83b}(1 - 3{,}62ae^{0{,}65b}), \qquad (37)$$

$$\lg k' = -0{,}514 + 0{,}31\lg a + 2{,}35\lg b. \qquad (38)$$

Выражения (37) и (38) позволяют находить не только значения параметров α_0 и k' по известным значениям параметров a и b, но также дают возможность решать обратную задачу.

Формулы (31) — (38) имеют ограниченные пределы применимости и справедливы при $b \approx 0{,}50 \div 0{,}60$, $k' \approx 0{,}013 \div 0{,}044$. Большинство текстов, составленных из ключевых слов, удовлетворяет этим условиям.

Для литературных текстов $b \approx 0{,}30 \div 0{,}40$. В этом случае кривые роста новых слов хорошо описываются формулой (при $10^4 < X < 10^8$)

$$Y = X^{\frac{1}{\sqrt{1+\alpha \ln X}}}, \qquad (39)$$

причем, параметр α в явном виде выражается через переменные Y и X:

$$\alpha = \frac{1}{\ln Y}\left(\frac{\ln X}{\ln Y} - \frac{\ln Y}{\ln X}\right).$$

Здесь величина α представляет собой среднее арифметическое из двух ее значений, которые можно найти из уравнения (18) при $u = 1/2$ и $u = -1$ и примерно соответствует величине $u = -1/4$.

Исследование уравнения (39) (с учетом равенства $\alpha = \alpha_0 + k \ln X$) показывает, что переменная Y достигает максимального значения при $X = \infty$, при этом $Y_{max} = e^{\sqrt{\frac{1}{k}}}$.

Формула (39) позволяет получить более простые выражения для определения полноты словаря и вероятности появления нового слова в тексте заданной длины, чем в случае использования формулы (32). Дифференцируя (39) по X (при $\alpha = \alpha_0 + k \ln X$), после преобразований получим

$$\frac{dY}{dX} = \frac{Y}{X}\left(\frac{\ln Y}{\ln X}\right)^3\left(1 + \frac{\alpha_0}{2} \ln X\right). \qquad (40)$$

Все приведенные выше формулы были получены для случайно составленной выборки, т. е. для такого условного текста, в котором разные слова появляются независимо и случайно. В реальном же тексте лексико-грамматические связи накладывают определенные ограничения на сочетаемость слов. Однако они проявляются в тексте слабо и не могут существенно изменить характер кривой роста новых слов в связном тексте по сравнению с соответствующей кривой в выборке. Поэтому формулы (32) и (39) могут быть использованы также для описания кривых роста новых слов в связных лексически однородных текстах.

Запишем уравнение прямой, характеризующей связный текст, в виде $\alpha_т = \alpha_{0т} + k_т \ln X$ в отличие от прямой $\alpha = \alpha_0 + k \ln X$ для случайной выборки. Здесь величина X обозначает длину текста и длину выборки.

Опытная проверка показала, что для литературных текстов значения параметра $\alpha_т$ близки к нулю. Например, для произведений А. С. Пушкина в случае использования формулы (39) величина $\alpha_{0т} = 0{,}0086$, для «Поднятой целины» М. А. Шолохова $\alpha_{0т} = -0{,}0009$. При этом параметр $k_т$ в обоих случаях примерно равен 0,0037. Принимая в формуле (40) $\alpha_0 = 0$, получаем возможность оценивать вероятность появления нового слова в тексте заданной длины лишь по двум известным из опыта величинам X и Y:

$$\frac{dY}{dX} \approx \frac{Y}{X}\left(\frac{\ln Y}{\ln X}\right)^3.$$

Расчеты, произведенные на основе опытных данных, показали, что параметр выборки k не зависит от величины X, а параметр α_0 с ростом X уменьшается пропорционально $\ln X$. В то же время параметры $\alpha_{0т}$ и $k_т$ не зависят от длины текста. При этом $k > k_т$, а отношение $k/k_т \approx 1{,}5$. Величина $\Delta k = k - k_т$ зависит от степени связности слов в тексте, а также от степени неоднородности текста по лексическому составу. В случае несвязного лексически однородного текста $\Delta k = 0$.

В заключение отметим, что параметры a и b распределения Вейбулла при $\Delta k > 0$ не являются постоянными. Параметр b практически не зависит от длины текста, а параметр a с ростом X уменьшается пропорционально $\ln X$.

ЛИТЕРАТУРА

1. Нешитой В. В. Законы распределения слов в тексте и его лексическая параметризация. Канд. дисс. АН БССР. Ин-т языкознания. Минск, 1973.
2. Белоногов Г. Г. О некоторых статистических закономерностях в русской письменной речи.— «Вопросы языкознания», 1962, № 1.
3. Калинин В. М. Функционалы, связанные с распределением Пуассона, и статистическая структура текста.— Труды математического института АН СССР, М.—Л, 1965, Т. 79.
4. Арапов М. В., Ефимова Е. Н., Шрейдер Ю. А. О смысле ранговых распределений.— «Научно-техническая информация. Серия 2», М., 1975.

Поступила в редакцию
12.VI 1974