ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2016 Управление, вычислительная техника и информатика № 3 (36)

УДК 519.21:81'32
DOI: 10.17223/19988605/36/5

В.В. Поддубный

О ВОЗМОЖНОСТИ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ЭВОЛЮЦИИ ПОЛИСЕМИИ ЗНАКОВ ЕСТЕСТВЕННОГО ЯЗЫКА С ПОМОЩЬЮ НЕСТАЦИОНАРНЫХ ПРОЦЕССОВ РОЖДЕНИЯ И ГИБЕЛИ

Работа выполнена при поддержке государственного задания Минобрнауки России № 1.511.2014/К и Российского гуманитарного научного фонда (проект № 14-14-70010).

Рассматривается возможность математического моделирования эволюции полисемии ансамбля знаков естественного языка с помощью нестационарных процессов рождения и гибели. Показано, что адекватной математической моделью развития полисемии ансамбля знаков может служить скрытая нестационарная модель процессов рождения и гибели значений языковых знаков. Получено условное распределение состояний такого процесса при экспоненциальных спадах интенсивностей процессов рождения и гибели. Предложен критерий идентификации скрытой модели, дана его реализация на примере словаря языка А.С. Пушкина.

Ключевые слова: неоднородный процесс рождения и гибели; скрытая марковская модель; идентификация модели; языковой знак; полисемия.

В работах [1−3] рассматривалась диссипативная стохастическая динамическая модель развития полисемии языковых знаков как детерминированная модель эволюции полисемии отдельного знака со случайными флуктуациями параметров от знака к знаку в статистическом ансамбле знаков. Модель показала хорошее согласие с экспериментально наблюдаемыми распределениями полисемии языковых знаков, полученными из представительных толковых словарей русского и английского языков. Однако детерминированный характер эволюции полисемии каждого отдельного знака представляется маловероятным. Более естественно предположить, что индивидуальная эволюция полисемии отдельного языкового знака является нестационарным стохастическим процессом.

В соответствии с современными лингвистическими представлениями о развитии полисемии языкового знака естественного языка [4] знак возникает в языке в некоторый случайный момент времени в некотором определённом (обычно единственном) смысловом значении. Затем к этому смысловому значению последовательно добавляются новые значения, как правило, всё более абстрактные. Процесс роста количества новых значений знака протекает с постоянным замедлением, пока не иссякнет способность знака к порождению новых значений. Одновременно с этим процессом (но, возможно, с некоторым запаздыванием) начинается процесс выхода из употребления первоначальных (наиболее конкретных) значений знака. Этот процесс протекает также с замедлением, но значительно медленнее процесса роста новых значений. Скорость выхода из употребления старых значений знака сначала меньше скорости роста числа новых значений, и количество не вышедших из употребления значений знака (его полисемия) сначала растёт. Но со временем скорость роста числа новых значений знака становится ниже скорости выпадения из употребления старых значений, и происходит обратный процесс – полисемия знака начинает убывать, пока не выйдет из употребления последнее значение знака, а с ним и сам знак. На этом жизненный цикл знака заканчивается. Кривая этого жизненного цикла, выражающая зависимость полисемии знака от времени, представляется унимодальной кривой с максимумом, смещённым к началу процесса развития полисемии знака.

Если предположить, что процессы появления и выпадения из употребления значений знака являются случайными марковскими, хотя, очевидно, нестационарными (неоднородными), представляется возможным использовать в качестве стохастической модели развития полисемии знака модель неоднородного марковского процесса рождения и гибели. В статистическом ансамбле знаков естественного языка параметры модели флуктуируют от знака к знаку с определёнными, но неизвестными распределениями вероятностей, в силу чего модель оказывается скрытой. Поставим задачу нахождения условного (с фиксированными значениями параметров) распределения вероятностей состояний неоднородного процесса рождения-гибели, а затем исследуем возможность оценки скрытых распределений вероятностей параметров ансамбля таких процессов, обеспечивающих максимальную близость теоретического распределения полисемии с наблюдаемым эмпирическим распределением, полученным из толкового словаря языка А.С. Пушкина.

1. Математическая модель неоднородного процесса рождения и гибели

1.1. Система уравнений Колмогорова

Составим систему дифференциальных уравнений Колмогорова, описывающих вероятностную динамику неоднородного марковского процесса рождения и гибели. Пусть очередной языковой знак появляется в языке в момент времени t_0 хотя бы в одном определённом смысловом значении. С этого момента начинается процесс рождения и гибели новых значений языкового знака вплоть до момента гибели последнего значения и выхода знака из употребления. Пусть $P_n(t)$ – вероятность того, что в момент времени $t \geq t_0$ знак имеет n значений. Если в начальный момент $n = n_0 \geq 1$, то $P_{n_0}(t_0) = 1$. Пусть $\lambda(t)$ – интенсивность процесса рождения новых значений в момент времени t, а $\mu(t)$ – интенсивность процесса гибели (выхода из употребления) уже имеющихся значений. Запишем незамкнутую систему дифференциальных уравнений Колмогорова, определяющую эволюцию вероятности числа живущих в момент времени t значений знака как неоднородного марковского процесса рождения и гибели значений:

$$\frac{dP_0(t)}{dt} = -\lambda(t)P_0(t) + \mu(t)P_1(t),$$

$$\frac{dP_n(t)}{dt} = \lambda(t)P_{n-1}(t) - (\lambda(t) + \mu(t))P_n(t) + \mu(t)P_{n+1}(t),\ P_n(t_0) = \delta_{n,n_0},\ n = 1,2,\ldots, \tag{1}$$

где $\delta_{n,n_0} = \begin{cases} 1, & n = n_0 \\ 0, & n \neq n_0 \end{cases}$ – символ Кронекера. При этом должно выполняться условие нормировки

$$\sum_{n=0}^{\infty} P_n(t) = 1.$$

1.2. Производящая функция

Для решения незамкнутой неавтономной (с переменными коэффициентами) системы дифференциальных уравнений Колмогорова (1) воспользуемся методом производящей функции, аналогично тому, как это делается в случае незамкнутой автономной системы (например, в [5. С. 287–291]):

$$f(t,s) = \sum_{n=0}^{\infty} P_n(t)s^n. \tag{2}$$

Зная производящую функцию $f(t, s)$, распределение $P_n(t)$ можно найти по формуле обращения

$$P_n(t) = \frac{1}{n!}\left.\frac{\partial^n f(t,s)}{\partial s^n}\right|_{s=0},\ n = 0,1,2,\ldots. \tag{3}$$

Действительно, разложив функцию $f(t, s)$ в ряд Маклорена, получим

$$f(t,s) = \sum_{n=0}^{\infty} \frac{1}{n!}\left.\frac{\partial^n f(t,s)}{\partial s^n}\right|_{s=0} s^n.$$

Сравнивая эту формулу с формулой (2), получим (3).

Перейдём от незамкнутой системы обыкновенных дифференциальных уравнений (1) для распределения $P_n(t)$ к дифференциальному уравнению в частных производных для производящей функции $f(t, s)$. Найдём частную производную

$$\frac{\partial f(t,s)}{\partial t}=\sum_{n=0}^{\infty}\frac{dP_n(t)}{dt}s^n ,$$

подставив в неё вместо производных $dP_n(t)/dt$ правые части уравнений (1). Принимая во внимание определение (2) производящей функции и вытекающее из этого определения равенство

$$\frac{\partial f(t,s)}{\partial s}=\sum_{n=0}^{\infty}P_n(t)ns^{n-1} ,$$

получим дифференциальное уравнение в частных производных первого порядка для производящей функции $f(t, s)$

$$\frac{\partial f(t,s)}{\partial t}=-\lambda(t)(1-s)f(t,s)+\mu(t)(1-s)\frac{\partial f(t,s)}{\partial s},\ f(t_0,s)=s^{n_0},\ n_0>0,\ t\ge t_0 . \tag{4}$$

Введя переменные $p=\partial f(t,s)/\partial t$ и $q=\partial f(t,s)/\partial s$, запишем уравнение (4) в виде

$$F(t,s,f,p,q)=-\lambda(t)(1-s)f+p-\mu(t)(1-s)q=0 . \tag{5}$$

Ему эквивалентна система обыкновенных дифференциальных уравнений для характеристик

$$\frac{dt}{F_p}=\frac{ds}{F_q}=\frac{df}{pF_p+qF_q}, \tag{6}$$

где $F_p=\partial F/\partial p=1$, $F_q=\partial F/\partial q=-\mu(t)(1-s)$, $pF_p+qF_q=p-\mu(t)(1-s)q=\lambda(t)(1-s)f$, причём последнее равенство записано с учётом равенства (5). Тогда система (6) примет вид

$$\mu(t)dt=-\frac{ds}{1-s},\ \lambda(t)ds=\mu(t)\frac{df}{f} .$$

Интегрируя каждое из уравнений, получаем

$$\int\mu(t)dt-\ln(1-s)=c_1,\ \lambda(t)s-\mu(t)\ln f=c_2 ,$$

где c_1, c_2 – произвольные постоянные интегрирования. Очевидно, c_2 можно рассматривать как произвольную функцию W от c_1: $c_2 = W(c_1)$, так что

$$\lambda(t)s-\mu(t)\ln f=W\left(\int\mu(t)dt-\ln(1-s)\right), \tag{7}$$

откуда

$$f(t,s)=\exp\left(\frac{1}{\mu(t)}\left(\lambda(t)s-W\left(\int_{t_0}^{t}\mu(t)dt-\ln(1-s)\right)\right)\right).$$

Очевидно, для существования производящей функции при любом t, в том числе при $t\to\infty$, необходимо, чтобы интенсивность процесса гибели нигде не обращалась в 0: $\mu(t)>0\quad\forall t\ge t_0$. При этом интенсивность процесса рождения может обращаться в 0 (например, при $t\to\infty$).

Для нахождения вида функции W воспользуемся (аналогично [5]) начальным условием $f(t_0,s)=s^{n_0}$. При $t = t_0$ равенство (7) примет вид

$$\lambda_0 s-\mu_0 n_0\ln s=W(-\ln(1-s)), \tag{8}$$

где $\lambda_0 = \lambda(t_0)$, $\mu_0 = \mu(t_0)$. Обозначив $y = -\ln(1 - s)$, получим $s = 1 - \exp(-y)$. Подставляя эти выражения в равенство (8), получим вид функции W: $W(y)=\lambda_0(1-\exp(-y))-n_0\mu_0\ln(1-\exp(-y))$. Следовательно, выражение для производящей функции принимает окончательный вид

$$f(t,s)=\left(1-(1-s)\exp\left(-\int_{t_0}^{t}\mu(t)dt\right)\right)^{n_0\mu_0/\mu(t)}\cdot\exp\left(\frac{1}{\mu(t)}\left(\lambda(t)s-\lambda_0\left(1-(1-s)\exp\left(-\int_{t_0}^{t}\mu(t)dt\right)\right)\right)\right). \tag{9}$$

1.3. Распределение вероятностей нестационарного процесса рождения и гибели

Для нахождения закона распределения вероятностей нестационарного процесса рождения и гибели воспользуемся формулой обращения (3). Для упрощения вида формулы (9) введём обозначения

$$a(t)=\frac{\mu_0}{\mu(t)},\ b(t)=\exp\left(-\int_{t_0}^{t}\mu(t)dt\right),\ c(t)=\frac{\lambda(t)-\lambda_0 b(t)}{\mu(t)} . \tag{10}$$

Тогда формула (9) примет вид

$$f(t,s)=\big((1-b(t))+b(t)s\big)^{n_0a(t)}\cdot\exp\left(-\frac{\lambda_0}{\mu(t)}(1-b(t))+c(t)s\right). \tag{11}$$

Обозначив

$$u(t,s)=\big((1-b(t))+b(t)s\big)^{n_0a(t)},\ v(t,s)=\exp(c(t)s), \tag{12}$$

ещё более упростим формулу (11), выделив множители, явно зависящие от переменной s:

$$f(t,s)=\exp\left(-\frac{\lambda_0}{\mu(t)}(1-b(t))\right)\cdot u(t,s)\cdot v(t,s). \tag{13}$$

Для вычисления вероятностей $P_n(t)$, $n=0,1,2,\ldots$, необходимо найти n-ю частную производную по s от этой функции в точке $s=0$. Поскольку, как видно из (13), эта функция пропорциональна произведению двух функций, зависящих от s, для вычисления производной воспользуемся известной формулой дифференцирования Лейбница

$$(uv)^{(n)}=\sum_{k=0}^{n}\binom{n}{k}u^{(k)}v^{(n-k)}. \tag{14}$$

Дифференцируя выражения (12), получаем

$$u(t,s)^{(k)}=(n_0a(t))(n_0a(t)-1)\cdots(n_0a(t)-k+1)b(t)^k\big((1-b(t))+b(t)s\big)^{n_0a(t)-k},$$

$$v(t,s)^{(n-k)}=c(t)^{n-k}\exp(c(t)s).$$

Учитывая, что

$$(n_0a(t))(n_0a(t)-1)\cdots(n_0a(t)-k+1)=\frac{\Gamma(n_0a(t)+1)}{\Gamma(n_0a(t)-k+1)},$$

получаем решение незамкнутой системы (1) дифференциальных уравнений Колмогорова

$$P_n(t)=\frac{1}{n!}\left.\frac{\partial^n f(t,s)}{\partial s^n}\right|_{s=0}=$$

$$=\exp\left(-\frac{\lambda_0}{\mu(t)}(1-b(t))\right)(1-b(t))^{n_0a(t)}\sum_{k=0}^{n}\frac{c(t)^{n-k}\,\Gamma(n_0a(t)+1)}{k!(n-k)!\Gamma(n_0a(t)-k+1)}\left(\frac{b(t)}{1-b(t)}\right)^k, \tag{15}$$

где $\Gamma(\cdot)$ – гамма-функция. Полученное распределение необходимо подчинить условию нормировки.

1.4. Частный случай: распределение вероятностей нестационарного процесса гибели

Частный случай процесса только гибели получается в отсутствие процесса рождения, когда $\lambda(t)\equiv 0$, а следовательно, когда $c(t)\equiv 0$. Распределение вероятностей такого процесса легко получить формально из общей формулы (15) при $\lambda_0=0$ и $c(t)\equiv 0$, когда в сумме по k остаётся только одно слагаемое – при $k=n$:

$$P_n(t)=\frac{\Gamma(n_0a(t)+1)}{n!\Gamma(n_0a(t)-n+1)}(1-b(t))^{n_0a(t)}\left(\frac{b(t)}{1-b(t)}\right)^n\cdot 1(n\le n_0),\ n=\overline{0,n_0},$$

где $1(n\le n_0)$ – индикатор условия, записанного в скобках (равен 1, если условие выполнено, и 0 в противном случае). Полученное распределение необходимо подчинить условию нормировки.

1.5. Частный случай: распределение вероятностей нестационарного процесса рождения

Частный случай процесса чистого рождения, когда $\mu(t)\equiv 0$, $a(t)\equiv 1$, $b(t)\equiv 1$, а $c(t)$ неограниченно возрастает, затруднительно получить из общего распределения (15), но легко получить, используя частный вид уравнения (4) для производящей функции при $\mu(t)\equiv 0$:

$$\frac{\partial f(t,s)}{\partial t}=-\lambda(t)(1-s)f(t,s),\ f(t_0,s)=s^{n_0},\ n_0>0,\ t\ge t_0. \tag{16}$$

Это уравнение при любом фиксированном s является обыкновенным дифференциальным уравнением первого порядка с разделяющимися переменными. Интегрируя его с заданным в (16) начальным условием, получаем

$$f(t,s)=s^{n_0}\exp\left(-(1-s)\int_{t_0}^{t}\lambda(t)dt\right).$$

Вычисление распределения $P_n(t)$ также производим по формуле обращения (3) с использованием обозначения

$$g(t)=-\int_{t_0}^{t}\lambda(t)dt,$$

представления

$$f(t,s)=\exp\left(-\int_{t_0}^{t}\lambda(t)dt\right)\cdot u(t,s)\cdot v(t,s)$$

и формулы Лейбница (14) для вычисления производных, где функции $u(t,s)$ и $v(t,s)$ имеют вид

$$u(t,s)=s^{n_0},\ v(t,s)=\exp(g(t)s).$$

Дифференцируя их по s, получаем

$$u^{(k)}=\begin{cases}n_0\,(n_0-1)\cdots(n_0-k+1)s^{n_0-k},\ k\le n_0,\\ 0,\ k>n_0\end{cases}\quad v^{(n-k)}=g(t)^{n-k}\exp(g(t)s).$$

Тогда при $s=0$ в сумме (14) остаётся только одно слагаемое при $k=n_0$ и $n\ge n_0$, и распределение принимает вид

$$P_n(t)=\frac{1}{n!}\frac{\partial^n f(t,s)}{\partial s^n}\bigg|_{s=0}=\frac{1}{(n-n_0)!}\left(\int_{t_0}^{t}\lambda(t)dt\right)^{n-n_0}\cdot\exp\left(-\int_{t_0}^{t}\lambda(t)dt\right)\cdot 1(n\ge n_0),\ n=\overline{n_0,\infty},\tag{17}$$

где $1(n\ge n_0)$ – индикатор условия, записанного в скобках. Формула (17) выражает распределение Пуассона для $n\ge n_0$, что хорошо известно для марковского процесса чистого рождения. Полученное распределение автоматически удовлетворяет условию нормировки.

1.6. Условие остановки неоднородного процесса рождения и гибели

Возвратимся к формуле (15), представляющей распределение вероятностей $P_n(t)$ состояний процесса рождения и гибели. Нетрудно видеть, что только входящая в него множителем функция $c(t)$, определяемая формулами (10), при некотором $t=t^*$ может обратиться в 0, вследствие чего $P_n(t^*)$ при всех $n>0$ обращается в 0, а $P_0(t^*)=1$. Следовательно, все ненулевые состояния в этот момент времени поглощаются и процесс рождения-гибели останавливается.

Рассмотрим подробнее условие остановки процесса. Выпишем функцию $c(t)$ из (10):

$$c(t)=\frac{\lambda(t)}{\mu(t)}-\frac{\lambda_0}{\mu(t)}\exp\left(-\int_{t_0}^{t}\mu(t)dt\right).\tag{18}$$

Предположим, что интенсивности процессов рождения и гибели монотонно уменьшаются с ростом t и не обращаются в 0 ни при каком конечном $t>t_0$. Пусть для определённости они спадают по экспоненциальному закону:

$$\lambda(t)=\lambda_0\exp(-(t-t_0)/\tau_1),\ \mu(t)=\mu_0\exp(-(t-t_0)/\tau_2),\tag{19}$$

где λ_0, μ_0 – начальные (в момент t_0) интенсивности, τ_1, τ_2 – постоянные времени спадов интенсивностей. Поскольку интенсивности (19) положительны при конечном $t\ge t_0$, функция $\varphi(t)=c(t)\mu(t)/\lambda_0$ имеет тот же знак, что и $c(t)$. Выпишем её с учётом (19):

$$\varphi(t)=\exp(-(t-t_0)/\tau_1)-\exp(-\mu_0\tau_2(1-\exp(-(t-t_0)/\tau_2))).\tag{20}$$

При $t=t_0$ эта функция обращается в 0, а её производная принимает значение $d\varphi(t_0)/dt=\mu_0-1/\tau_1$. С ростом t функция $\varphi(t)$ (и, следовательно, $c(t)$) либо становится всюду отрицательной (при $\mu_0\tau_1\le 1$), что недопустимо для существования (неотрицательности) распределения вероятностей ненулевых значений n, либо (при $\mu_0\tau_1 > 1$) возрастает, достигает положительного максимума в некоторой точке $t_{max} > t_0$, а затем спадает до значения 0 в некоторой точке $t^* > t_{max}$ и далее уходит в отрицательную область, принимая отрицательное значение $-\exp(-\mu_0\tau_2)$ при $t \to \infty$. В этом случае уравнение $\varphi(t) = 0$ имеет корень t^*, являющийся точкой остановки процесса рождения-гибели с вероятностью 1. Таким образом, ненулевое состояние процесса рождения-гибели с экспоненциально спадающими интенсивностями возможно только при $\mu_0\tau_1 > 1$ и только в интервале времени от $t = t_0$ до $t = t^*$, так что длительность жизни T процесса рождения-гибели не превышает разности $t^* - t_0$. Такой процесс (с ограниченным временем жизни) будем называть финитным.

На рис. 1 в качестве примера представлено семейство кривых $\varphi(t)$ при $\tau_1 = 0{,}4286$, $\tau_2 = 0{,}1429$ и $\mu_0\tau_1$, изменяющемся с шагом 0,5 в интервале от 0 до 2,5.

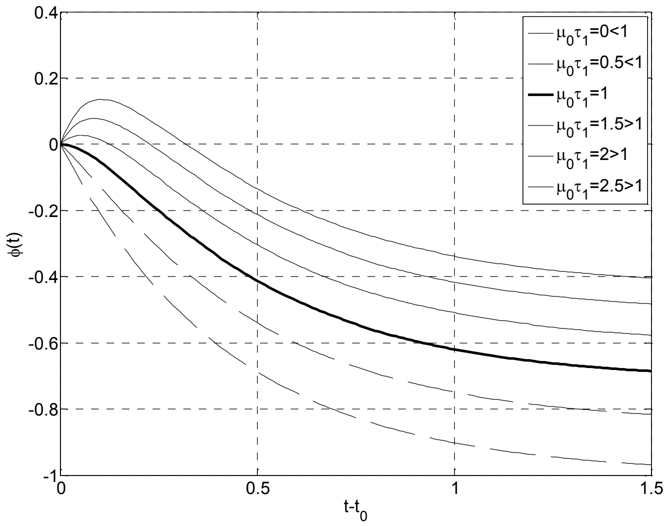

Рис. 1. Функция $\varphi(t)$

Заметим, что эффект остановки процесса рождения-гибели с вероятностью 1 не имеет места для однородного процесса, когда интенсивности постоянны (равны λ_0, μ_0), потому что для однородного процесса

$$\varphi(t) = 1 - \exp(-\mu_0(t - t_0)) > 0$$

при любых конечных $t > t_0$, так что уравнение $\varphi(t) = 0$ корней не имеет. Процесс останавливается только при случайном достижении состояния 0, но при этом $P_0(t^*) \neq 1$. Таким образом, однородный процесс рождения-гибели не является финитным.

2. Математическая модель статистического ансамбля неоднородных процессов рождения и гибели с монотонно убывающими интенсивностями

Рассмотрим теперь статистический ансамбль неоднородных процессов рождения и гибели. Ансамбль характеризуется случайными моментами t_0 возникновения каждого процесса рождения-гибели, а каждый из процессов рождения-гибели – случайными значениями параметров интенсивностей потоков рождения и гибели. Будем в дальнейшем предполагать, что интенсивности процессов рождения и гибели монотонно уменьшаются со временем t по экспоненциальному закону (19) от начальных значений λ_0 и μ_0 в момент времени $t = t_0$ до нуля при $t \to \infty$ с постоянными времени τ_1 и τ_2 соответственно. Тогда каждый процесс рождения-гибели в ансамбле будет характеризоваться условным распределением вероятностей (15) с пятью случайными параметрами t_0, λ_0, μ_0, τ_1, τ_2. Распределение вероятностей состояний ансамбля таких процессов рождения и гибели в каждый момент времени t получается усреднением выражения (15) по распределениям указанных пяти параметров:

$$P_n(t) = \int_{-\infty}^{t} dt_0 \int_0^{\infty} d\lambda_0 \int_0^{\infty} d\mu_0 \int_0^{\infty} d\tau_1 \int_0^{\infty} d\tau_2 P_n(t \mid t_0, \lambda_0, \mu_0, \tau_1, \tau_2) p(t_0, \lambda_0, \mu_0, \tau_1, \tau_2),$$

где $P_n(t \mid t_0, \lambda_0, \mu_0, \tau_1, \tau_2)$ представляется формулой (19), а $p(t_0, \lambda_0, \mu_0, \tau_1, \tau_2)$ – плотность совместного распределения вероятностей параметров t_0, λ_0, μ_0, τ_1, τ_2.

Предположим, что моменты t_0 возникновения событий, порождающих процессы рождения-гибели, образуют однородный пуассоновский поток независимых редких событий. Тогда параметр t_0 в бесконечном ансамбле таких процессов будет распределён на полуоси $(-\infty, t)$ равномерно. Естественно считать его статистически независимым от остальных параметров. Остальные четыре параметра λ_0, μ_0, τ_1, τ_2 также можно принять статистически независимыми. Однако при некоторых соотношениях между этими параметрами ненулевые состояния процесса рождения-гибели могут оказаться невозможными.

Во-первых, для ненулевой вероятности ненулевого состояния процесса рождения-гибели необходимо, чтобы в момент времени t была положительной функция $c(t)$, определяемая выражением (18) и входящая множителем в выражение (15) для функции распределения состояния процесса рождения-гибели. Следовательно, должна быть положительной функция $\varphi(t)$, определяемая выражением (20) при экспоненциальных спадах (19) интенсивностей процессов рождения и гибели. Как видно из анализа поведения во времени функции $\varphi(t)$ (рис. 1), для этого требуется выполнение неравенства

$$\mu_0 \tau_1 > 1. \tag{21}$$

Во-вторых, для финитного процесса рождения-гибели с экспоненциально убывающими интенсивностями полное (за всё время жизни процесса) среднее число $G_1(\infty)$ событий рождения и полное среднее число $G_2(\infty)$ событий гибели являются конечными. Поскольку в финитном процессе рождения-гибели ненулевые состояния с вероятностью 1 поглощаются за конечное время его жизни, естественно потребовать равенство этих средних:

$$G = G_1(\infty) = G_2(\infty),\ G_1(\infty) = \int_{t_0}^{\infty} \lambda(t) dt = \lambda_0 \tau_1,\ G_2(\infty) = \int_{t_0}^{\infty} \mu(t) dt = \mu_0 \tau_2,\ \lambda_0 \tau_1 = \mu_0 \tau_2 = G. \tag{22}$$

Получили два уравнения связей, позволяющих исключить переменные τ_1, τ_2 через переменные λ_0, μ_0 и новую переменную G:

$$\tau_1 = G/\lambda_0,\ \tau_2 = G/\mu_0. \tag{23}$$

Тогда неравенство (21) примет вид ограничения на переменную G:

$$G > \lambda_0/\mu_0. \tag{24}$$

Это значит, что при нарушении этого неравенства ненулевые состояния процесса рождения-гибели становятся невозможными.

В-третьих, чтобы разность процессов рождения и гибели с учётом (22) и (23) была в среднем неотрицательной, необходимо, чтобы

$$\lambda_0 > \mu_0. \tag{25}$$

Это условие можно проиллюстрировать графически. На рис. 2 представлены изменения во времени среднего накопленного к моменту t числа $G_1(t)$, $G_2(t)$ событий процессов рождения и гибели,

$$G_1(t) = \int_{t_0}^{t} \lambda(t) dt = \lambda_0 \tau_1 \left(1 - \exp\left(-\frac{t - t_0}{\tau_1}\right)\right),\ G_2(t) = \int_{t_0}^{t} \mu(t) dt = \mu_0 \tau_2 \left(1 - \exp\left(-\frac{t - t_0}{\tau_2}\right)\right),$$

а также их разности $G_1(t) - G_2(t)$ при выполнении условий (22) и соотношений (23).

Видно, что разность $G_1(t) - G_2(t)$, выражающая среднее состояние процесса рождения-гибели (среднее число «живущих» событий), при $\lambda_0 > \mu_0$ сначала быстро возрастает, достигает максимума, а затем медленно уменьшается, оставаясь неотрицательной величиной. Если бы неравенство было противоположным, разность стала бы отрицательной, а это невозможно, так как означало бы, что среднее число погибших элементов потока рождения-гибели превышает среднее число рождённых элементов. Следовательно, при нарушении неравенства (25) ненулевые состояния процесса рождения-гибели становятся невозможными.

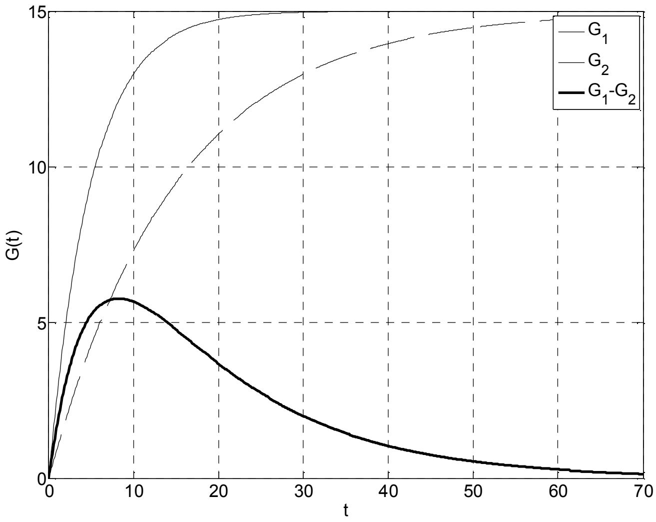

Рис. 2. Динамика среднего состояния процесса рождения-гибели при $G = 15$, $\lambda_0 = 3$, $\mu_0 = 1$ ($\lambda_0 > \mu_0$)

Таким образом, вместо четырёх параметров λ_0, μ_0, τ_1, τ_2 при нахождении безусловного распределения состояний процесса рождения-гибели можно обойтись тремя: λ_0, μ_0, G. С учётом естественной неотрицательности параметров λ_0, μ_0, G ограничения (24)–(25) определяют область возможных значений этих параметров при усреднении условного распределения:

$$P_n(t) = \int_{-\infty}^{t} p(t_0)dt_0 \int_0^{\infty} d\lambda_0 \int_0^{\lambda_0} d\mu_0 \int_{\lambda_0/\mu_0}^{\infty} dG \cdot P_n(t \mid t_0, \lambda_0, \mu_0, G)p(\lambda_0, \mu_0, G). \qquad (26)$$

Условное распределение $P_n(t \mid t_0, \lambda_0, \mu_0, G)$ представляется выражением (15) с входящими в него функциями $a(t)$, $b(t)$, $c(t)$, определяемыми выражениями (10) с учётом (19) и (23).

3. Математическая модель скрытого марковского процесса рождения и гибели и её идентификация

Статистический ансамбль неоднородных марковских процессов рождения-гибели со случайными параметрами при неизвестных распределениях параметров представляется скрытым марковским процессом рождения-гибели. Этот процесс наблюдаем, тогда как его параметры являются ненаблюдаемыми случайными величинами. Возникает вопрос, при каких распределениях параметров наблюдаемый процесс рождения-гибели имеет теоретическое распределение вероятностей состояний, максимально близкое к эмпирическому распределению?

Задача отыскания наилучшей статистической оценки распределения $p(\lambda_0, \mu_0, G)$ по наблюдаемому эмпирическому распределению $\{P_{nэ}(t), n = 1,2,\ldots,N\}$, где N – максимальное наблюдаемое в эмпирическом распределении значение n, является задачей статистической идентификации наблюдаемого скрытого процесса рождения-гибели и сводится к минимизации по $p(\lambda_0, \mu_0, G)$ расхождения между теоретическим распределением (26) с ядром (15) и эмпирическим распределением. Для корректного решения этой задачи можно использовать известные методы тихоновской регуляризации.

В качестве критерия идентификации (критерий близости распределений) целесообразно выбрать логарифмический среднеквадратический критерий вида

$$J = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\log P_n(t) - \log P_{nэ}(t)}{\log P_{nэ}(t)}\right)^2 \Rightarrow \min_{p(\lambda_0, \mu_0, G)}. \qquad (27)$$

Логарифмическая форма критерия удобна в случае больших (на несколько порядков) различий значений фигурирующих в критерии распределений при разных n.

Минимизация (27) с вычислением многомерного интеграла (26) представляет определённые вычислительные трудности, связанные, прежде всего, с преодолением некорректности и большим объёмом вычислений. Уменьшить число вычислений можно, заменяя интегралы суммами со сравнительно небольшими (приемлемыми с вычислительной точки зрения) числами слагаемых. При этом, естественно, снижается точность вычислений. Опуская детали вычислительной схемы, приведём результаты вычислений оптимальных значений теоретической функции распределения $P_{n\,\mathrm{opt}}(t)$, максимально приближенной к эмпирическому распределению $P_{nэ}(t)$ по критерию (27).

4. Идентификация математической модели скрытого неоднородного марковского процесса рождения и гибели по эмпирическому распределению полисемии языка А.С. Пушкина

В качестве эмпирического распределения $P_{nэ}$ возьмём распределение $P_{n\,\mathrm{Pushkin}}$ полисемии слов языка А.С. Пушкина [6]. В двойном логарифмическом масштабе это распределение представлено на рис. 3 тонкой кривой. Полужирной кривой показано оптимальное распределение $P_{n\,\mathrm{opt}}(t)$, вычисленное с использованием критерия (27) для некоторого фиксированного момента времени t без усреднения по t_0 в (26) (t_0 взято равным 0). Диапазоны значений параметров, на которых вычислялись их распределения: G – от 10 до 20 с шагом 0,5; λ_0 – от 0,1 до 6,1 с шагом 0,5; μ_0 – от 0,1 до 5,1 с шагом 0,5. Из рис. 3 видно хорошее согласие теоретического распределения с эмпирическим (достигнутый уровень значимости $p = 0{,}9971$ по критерию Колмогорова–Смирнова), что свидетельствует о возможности моделирования процесса развития полисемии языковых знаков скрытым марковским процессом рождения-гибели.

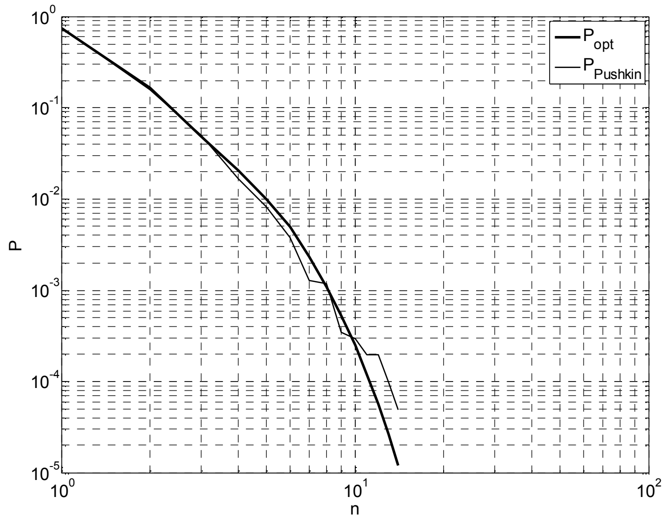

Рис. 3. Теоретическое распределение вероятностей состояний $P_{n\,\mathrm{opt}}$ неоднородного процесса рождения и гибели и эмпирическое распределение вероятностей $P_{n\,\mathrm{Pushkin}}$ значений полисемии языка А.С. Пушкина

Заключение

В работе выдвинута и подтверждена экспериментальными данными гипотеза о возможности математического моделирования процессов развития полисемии знаков естественного языка скрытыми нестационарными финитными марковскими моделями рождения и гибели. Получена аналитическая форма условного распределения вероятностей такого процесса при экспоненциально спадающих интенсивностях процессов рождения и гибели. Предложен критерий идентификации скрытой модели. Проведено приближённое численное решение задачи идентификации модели и вычислено безусловное одномоментное теоретическое распределение полисемии, соответствующее эмпирическому распределению полисемии языковых знаков словаря А.С. Пушкина. Получено хорошее согласие теоретического и экспериментального распределений полисемии.

ЛИТЕРАТУРА

1. Поддубный В.В., Поликарпов А.А. Диссипативная стохастическая динамическая модель развития языковых знаков // Компьютерные исследования и моделирование. 2011. Т. 3, № 2. С. 103–124.

2. Poddubny V.V., Polikarpov A.A. Stochastic Dynamic Model of Evolution of Language Sign Ensembles // Methods and Applications of Quantitative Linguistics. Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO) / ed. by Ivan Obradović, Emmerich Kelih and Reinhard Kohler. Belgrade, 2013. P. 69–83.
3. Poddubnyy V., Polikarpov A. Evolutionary Derivation of Laws for Polysemic and Age-Polysemic Distributions of Language Sign Ensembles // Recent Contributions to Quantitative Linguistics / ed. by A. Tuzzi, M. Benešová, J. Macutek. Walter de Gruyter GmbH, 2015. P. 115–124.
4. Поликарпов А.А. Модель жизненного цикла знака: К теоретическим основаниям исторической лексикологии и дериватологии // Славянская лексикография / ред. М.И. Чернышева. М. : Азбуковник, 2013. С. 679–702.
5. Fisz M. Probability Theory and Mathematical Statistics. New York ; London ; Sydney : John Wiley & Sons, 1967. 680 p.
6. Словарь языка Пушкина : в 4 т. 2-е изд., доп. / отв. ред. В.В. Виноградов ; Российская академия наук. Ин-т рус. яз. им. В.В. Виноградова. М. : Азбуковник, 2000.

Поддубный Василий Васильевич, д-р техн. наук, профессор. E-mail: vvpoddubny@gmail.com
Томский государственный университет

Поступила в редакцию 1 апреля 2016 г.

Poddubny Vasiliy V. (Tomsk State University, Russian Federation).

On the possibility of mathematical modelling of the evolution of the polysemy of natural language signs with using of non-stationary birth-death processes.

Keywords: heterogeneous process of birth and death; hidden Markov model; model identification; language sign; polysemy.

DOI: 10.17223/19988605/36/5

We consider the possibility of mathematical modeling of the evolution of polysemy of ensemble of signs of natural language by means of non-stationary processes of birth and death. We showed that an adequate mathematical model of polysemy of ensemble of signs might be built on the base of hidden non-stationary model of the birth and death processes of the meanings of linguistic signs. We assume exponential decay of the intensities of the processes of birth and death:

$$\lambda(t)=\lambda_0 \exp(-(t-t_0)/\tau_1),\ \mu(t)=\mu_0 \exp(-(t-t_0)/\tau_2),$$

where t is the current time; t_0 is the time moment when the sign appears in the ensemble; λ_0, μ_0 are the initial values of intensities of the processes of birth and death; $\tau_1 = G / \lambda_0$, $\tau_2 = G / \mu_0$ are time decay constants of intensities, and G is the average number of meanings, which the sign may birth and lose during his life:

$$G=\int_{t_0}^{\infty}\lambda(t)dt=\lambda_0\tau_1\ ,\ G=\int_{t_0}^{\infty}\mu(t)dt=\mu_0\tau_2\ .$$

We received the conditional (with fixed parameters t_0, λ_0, μ_0, G) probability distribution of states n of this process:

$$P_n(t\,|\,\vartheta)=\exp\left(-\frac{\lambda_0}{\mu(t)}\bigl(1-b(t)\bigr)\right)\bigl(1-b(t)\bigr)^{n_0a(t)}\sum_{k=0}^{n}\frac{c(t)^{n-k}\,\Gamma\bigl(n_0a(t)+1\bigr)}{k!(n-k)!\Gamma\bigl(n_0a(t)-k+1\bigr)}\left(\frac{b(t)}{1-b(t)}\right)^k,$$

where

$$a(t)=\frac{\mu_0}{\mu(t)},\ b(t)=\exp\left(-\int_{t_0}^{t}\mu(t)dt\right),\ c(t)=\frac{\lambda(t)-\lambda_0 b(t)}{\mu(t)}.$$

In the hidden model of the statistical ensemble of processes of birth and death the parameters t_0, λ_0, μ_0, G of each individual process (of each linguistic sign) randomly vary in relation of each to other, subject to certain distribution laws. Under the assumption of a Poisson distribution of the flow of signs, the distribution density of the parameter t_0 can be considered as uniform on a large enough time interval, while the distributions of parameters λ_0, μ_0, G are unknown. Unconditional probability distribution $P_n(t)$ of the state n of an ensemble of the processes of birth-death (of the polysemy of an ensemble of signs) at moment t is the mathematical expectation of the conditional distribution $P_n(t|\theta)$ over the distribution of parameters t_0, λ_0, μ_0, G.

We have solved the task of estimation of the parameter distributions (for identifying of hidden model) according to the empirical polysemy distribution P_{ne} obtained from a representative dictionary, with the subsequent calculation of the optimal theoretical distribution $P_n(t)$. As an identification criterion (criterion of proximity of distribution), we select a logarithmic RMS criterion of type:

$$J=\frac{1}{n_0}\sum_{n=1}^{n_0}\left(\frac{\log P_n(t)-\log P_{nэ}(t)}{\log P_{nэ}(t)}\right)^2\Rightarrow\min_{p(\lambda_0,\mu_0,G)},$$

convenient for large (several orders of magnitude) changes in distributions for different n. The criterion was implemented on example of using of the dictionary of Pushkin's language. We obtain a good agreement of distributions $P_n(t)$ and P_{ne} that confirms the possibility of using of hidden mathematical model of non-stationary process of birth-death for the simulation of polysemy evolution of the ensemble of signs of natural language.

REFERENCES

1. Poddubnyy, V.V. & Polikarpov, A.A. (2011) Dissipative Stochastic Dynamic Model of Language Signs Evolution. *Komp'yuternye issledovaniya i modelirovanie – Computer Research and Modeling*. 3(2). pp.103-124. (In Russian).
2. Poddubny, V.V. & Polikarpov, A.A. (2013) Stochastic Dynamic Model of Evolution of Language Sign Ensembles. *Methods and Applications of Quantitative Linguistics. Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)*. Belgrade. pp. 6983.
3. Poddubnyy, V. & Polikarpov, A. (2015) Evolutionary Derivation of Laws for Polysemic and Age-Polysemic Distributions of Language Sign Ensembles. In: Tuzzi, A., Benešová, M. & Macutek, J. (eds) *Recent Contributions to Quantitative Linguistics*. GmbH: Walter de Gruyter. pp. 115-124.
4. Polikarpov, A.A. (2013) Model' zhiznennogo tsikla znaka: K teoreticheskim osnovaniyam istoricheskoy leksikologii i derivatologii [Model of the Sign Life Cycle: To the Theoretical Foundations of Historical Lexicology and Word Formation]. In: Chernysheva, M.I. (ed.) *Slavyanskaya leksikografiya* [Slavic Lexicography]. Moscow: Azbukovnik. pp. 679-702.
5. Fisz, M. (1967) *Probability Theory and Mathematical Statistics*. 3rd ed. New York-London-Sydney: John Wiley & Sons.
6. Vinogradov, V.V. (ed.). (2000) *Slovar' yazyka Pushkina: v 4 t.* [Dictionary of Pushkin’s Language: in 4 vols]. 2nd ed. Moscow: Academy of Sciences of the USSR, Azbukovnyk.