
XIV ЭМ'2010

Российская академия наук
Министерство образования и науки РФ
Красноярский государственный торгово-экономический институт
Институт математики, Сибирский федеральный университет
Институт вычислительного моделирования СО РАН
Сибирский институт бизнеса, управления и психологии

ТРУДЫ

XIV МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО ЭВЕНТОЛОГИЧЕСКОЙ МАТЕМАТИКЕ И СМЕЖНЫМ ВОПРОСАМ

Красноярск
Красноярский государственный торгово-экономический институт
Сибирский федеральный университет

2010

УДК 519.248: [004.8+33+301+159.9]
Т 78

Труды XIV международной ЭМ'2010 конференции. Под ред. Олега Воробьёва. — Красноярск: Крас. гос. торг.-эконом. ин-т, Сиб. фед. ун-т, 2010. — 236 с.

ISBN 978-5-98153-165-1

ISBN 978-5-98153-165-1

Диссипативная стохастическая динамическая модель эволюции языковых знаков

Василий Васильевич Поддубный
Томский государственный университет
Факультет информатики
Томск
pvv@inet.tsu.ru

Анатолий Анатольевич Поликарпов
Московский государственный университет им. М.В. Ломоносова
Филологический факультет
Москва
polikarp@philol.msu.ru

Аннотация. *Предлагается диссипативная стохастическая динамическая модель эволюции языковых знаков, удовлетворяющая принципу "наименьшего действия", одному из фундаментальных вариационных принципов природы. Модель предполагает пуассоновский характер потока рождения языковых знаков, экспоненциальное (показательное) распределение ассоциативно-семантического потенциала (АСП) знака и оперирует разностными стохастическими уравнениями специального вида, вытекающими из принципа наименьшего действия для диссипативных процессов. Получаемые из модели распределения полисемии языковых знаков статистически значимо (по критерию Колмогорова-Смирнова) не отличаются от эмпирических распределений, полученных из 5 представительных словарей русского и английского языков.*

Ключевые слова. *Языковой знак, эволюция, ассоциативно-семантический потенциал, значение знака, полисемия, диссипативная стохастическая динамическая модель*

1 Введение

Согласно современным представлениям о развитии (эволюции) жизненного цикла языкового знака [2], любой языковой знак (слово) после своего появления в языке в некотором начальном значении может либо сохранять это значение в течение всей своей жизни, либо претерпевать эволюцию, последовательно рождая новые значения, все более и более абстрактные по смыслу, пока не будет полностью израсходован его так называемый ассоциативно-семантический потенциал (АСП) – способность порождать новые значения (своя для каждого знака). При этом скорость рождения новых значений на начальном участке процесса эволюции знака максимальна (хотя и своя для каждого знака), а затем постепенно падает до нуля. Через некоторое время τ_0 (свое для каждого знака) начинается аналогичный процесс выпадения из употребления сначала наименее абстрактных (наиболее конкретных) значений знака, а затем все более абстрактных, пока все значения знака не выйдут из употребления. При этом начальный участок процесса выпадения из употребления значений знака характеризуется наибольшей скоростью, которая затем постепенно уменьшается до нуля. Весь процесс выпадения значений знака из употребления идет более медленно, чем процесс рождения новых значений. В результате в каждый данный момент текущего времени после момента возникновения знака число актуальных (живущих) значений знака (его полисемия) сначала растет, достигает максимума, а затем постепенно падает. Через некоторое время (время жизни знака) это число становится равным нулю – знак исчезает из употребления. Кривая развития этого процесса во времени (кривая жизненного цикла языкового знака) – унимодальная кривая с максимумом, смещенным влево, к началу процесса.

Возникает вопрос, какой математической моделью может быть описан процесс развития языкового знака?

2 Детерминированная диссипативная динамическая модель эволюции языкового знака

2.1 Модель процесса рождения новых значений знака

Введем переменные, характеризующие процесс роста числа значений отдельного языкового знака. Обозначим через G максимальный АСП знака; через k – номер значения, появившегося на k-м шаге эволюции значений знака, $k = 1, 2, \ldots, G$; через t_k – момент появления k-го значения ($t_1 = 0$ – начальный момент, момент появления знака в его начальном значении); через v_k – скорость роста числа значений знака на k-м шаге эволюции (v_1 – начальная скорость).

Очевидно,

$$v_k = ((k+1) - k)/(t_{k+1} - t_k) = 1/\Delta t_k, \qquad (1)$$

где $\Delta t_k = t_{k+1} - t_k$ – промежуток времени между

рождениями $k+1$-го и k-го значений знака. В процессе рождения новых значений знака его АСП растрачивается, уменьшаясь на единицу при каждом рождении нового значения, так что на k-м шаге (уровне) эволюции АСП знака оказывается равным $G-k$ (после рождения k-го значения знака может родиться только $G-k$ новых значений).

Естественно предположить, что скорость рождения новых значений знака v_k пропорциональна АСП знака на k-м уровне эволюции:

$$v_k = a(G-k),\ k = \overline{1,G}. \qquad (2)$$

Тогда максимальная скорость роста равна $v_1 = a(G-1)$, а минимальная – нулю: $v_G = a(G-G) = 0$. Из соотношений (1)–(2) получаем рекуррентную формулу для моментов появления новых значений знака:

$$t_{k+1} = t_k + 1/(a(G-k)),\ k = \overline{1,G-1},\ t_1 = 0. \qquad (3)$$

Отсюда видно, что интервал времени $\Delta t_k = t_{k+1} - t_k$ обратно пропорционален АСП k-го значения знака, причем, чем больше величина коэффициента пропорциональности $1/a$, тем длиннее этот интервал. Следовательно, коэффициент $1/a$ имеет смысл некоторой "постоянной времени" $\tau = 1/a$ роста числа новых значений знака (чем больше τ, тем медленнее рост, т.к. больше Δt_k, и наоборот). Поэтому вместо коэффициента a более резонно использовать обратную величину τ, так что рекуррентное соотношение (3) примет вид:

$$t_{k+1} = t_k + \tau/(G-k),\ k = \overline{1,G-1},\ t_1 = 0. \qquad (4)$$

2.2 Процесс рождения новых значений знака как диссипативный процесс, удовлетворяющий принципу наименьшего действия

К сожалению, проанализировать характер поведения решения рекуррентного уравнения (4) затруднительно. Однако, если перейти временно к непрерывной переменной k, то рекуррентное уравнение (2) с учетом (1) и обозначения $\tau = 1/a$ можно записать в виде дифференциального уравнения:

$$dk(t)/dt = (G - k(t))/\tau,\ k(0) = 1. \qquad (5)$$

Это линейное дифференциальное уравнение с разделяющимися переменными. Его решение имеет вид:

$$k(t) = G - (G-1)\exp(-t/\tau),\ t \geq 0. \qquad (6)$$

Как видим, $k(t)$ экспоненциально растет с ростом t от значения $k(0) = 1$ до $k(\infty) = G$. Очевидно,

$$v(t) = dk(t)/dt = (G - k(t))/\tau, \qquad (7)$$

$$d^2k(t)/dt^2 = -(1/\tau)dk(t)/dt = -(1/\tau)v(t). \qquad (8)$$

Последнее равенство означает, что "сила инерции" $d^2k(t)/dt^2$ процесса роста числа значений знака в каждый момент текущего времени t уравновешивается "силой вязкого трения" $-(1/\tau)v(t)$. Следовательно, процесс роста новых значений знака является диссипативным процессом [1]. Первоначальная "кинетическая энергия" такого процесса, равная $T(0) = v^2(0)/2 = (G-1)^2/(2\tau^2)$, растрачивается на преодоление "сил вязкого трения", так что к моменту времени $t > 0$ ее остается

$$T(t) = v^2(t)/2 = (G - k(t))^2/(2\tau^2) < T(0). \qquad (9)$$

С другой стороны, этот остаток "кинетической энергии" способен совершить "работу" против "сил вязкого трения", равную

$$U(t) = -\int_{J_{k(t)}}^{G} (1/\tau)v(t)dk(t) = (G-k(t))^2/(2\tau^2). \qquad (10)$$

Эта "работа" называется "потенциальной энергией" (потенциальной функцией). Как видим, "потенциальная энергия" знака $U(t)$ на каждом уровне $k(t)$, т.е. в каждый момент времени t, равна "кинетической энергии" знака $T(t)$, $U(t) = T(t)$, так что его "функция действия" $S(t) = |T(t) - U(t)| \equiv 0$ и соответственно "функционал действия"

$$S = \int_0^{\infty} |T - U|dt = 0$$

принимают минимальные (именно, равные нулю) значения. Следовательно, процесс $k(t)$ роста числа новых значений знака в рассматриваемой математической модели подчиняется принципу "наименьшего действия" [4, 5], одному из фундаментальных вариационных принципов природы, впервые сформулированному Пьером Мопертюи в 40-х годах 18 века.

Частная производная $\partial U/\partial k(t) = -(1/\tau)v(t)$ от потенциальной функции U по координате $k(t)$ определяет "силу вязкого трения", "силу сопротивления движению", что характерно для диссипативных систем. Применительно к языковому знаку можно сказать, что она определяет силу сопротивления языковой среды рождению новых значений знака. Возвращаясь к дискретной модели роста числа новых значений знака, можно сказать, что процесс эволюции языкового знака есть дискретный аналог диссипативного процесса.

2.3 Модель процесса выхода из употребления значений знака и общая детерминированная модель эволюции знака

Естественно предположить, что аналогично, но более медленно протекает процесс выхода из употребления значений знака. Тогда, снабдив верхними индексами процесс рождения новых значений (индекс 1) и процесс выпадения значений из употребления (индекс 2), для i-го знака получим из (4):

$$t_{i,k+1}^{(1,2)} = t_{i,k}^{(1,2)} + \tau_i^{(1,2)}/(G_i - k),\ k = \overline{1, G_i - 1},$$
$$t_{i,1}^{(1)} = t_i,\ t_{i,1}^{(2)} = \tau_{0i} + t_i,\ \tau_i^{(2)} > \tau_i^{(1)},\ i = \overline{1, N}, \quad (11)$$

где t_i – момент появления в языке i-го знака (слова), N – число слов в языке. Очевидно, $L_{i,k} = t_{i,k}^{(2)} - t_{i,k}^{(1)}$ – длительность жизни k-го значения i-го знака. Нетрудно видеть, что эта величина подчиняется рекуррентному соотношению:

$$L_{i,k+1} = L_{i,k} + (\tau_i^{(2)} - \tau_i^{(1)})/(G_i - k),\ L_{i,1} = \tau_{0i},$$
$$k = \overline{1, G_i - 1},\ i = \overline{1, N}, \quad (12)$$

так что $L_{i,k+1} > L_{i,k}$, т.е. длительность жизни каждого значения любого i-го знака увеличивается с ростом k.

Полисемия i-го знака развивается с момента $t_{i,1}^{(1)} = t_i$ появления i-го знака в языке до момента $t_{i,G_i}^{(2)} = t_i + \tau_{0i} + L_{i,G_i}$ выхода из употребления последнего (G_i-го) значения i-го знака. Интервал времени длиной $L_i = \tau_{0i} + L_{i,G_i}$ от $t = t_{i,1}^{(1)}$ до $t = t_{i,G_i}^{(2)}$ – интервал жизненного цикла i-го знака.

Если на заданный момент времени T (на данном временном сечении) i-ый знак существует, т.е. уже появился, но еще не вышел из употребления, так что $t_i \le T < t_{i,G_i}^{(2)}$, то разность $A_i = T - t_i$ определяет возраст i-го знака в момент времени T, а все множество "живущих" на этот момент знаков с их возрастами определяет одномоментное распределение знаков по возрастам.

Модель (11)–(12) является детерминированной дискретной диссипативной динамической моделью развития жизненного цикла языкового знака.

Очевидно, параметры жизненного цикла i-го знака (их 5 в нашей детерминированной модели) t_i, G_i, τ_{0i}, $\tau_i^{(1)}$, $\tau_i^{(2)}$ различны для различных знаков. Этими параметрами определяются и моменты рождения $t_{i,k}^{(1)}$, и моменты выхода из употребления $t_{i,k}^{(2)}$, и длительности жизни $L_{i,k}$ каждого k-го значения i-го знака, и, наконец, само число G_i различных значений i-го знака (его АСП). Параметры $\tau_i^{(1)}$ и G_i определяют активность i-го знака, параметры τ_0 и L_{i,G_i} (т.е. τ_{0i}, $\tau_i^{(2)} - \tau_i^{(1)}$, G_i) – его стабильность.

3 Стохастическая диссипативная динамическая модель эволюции языковых знаков

Поскольку параметры модели характеризуют каждый знак, а все знаки разные, можно считать эти параметры для случайно выбранного знака случайными, подчиняющимися некоторым устойчивым законам распределения (в стационарном режиме появления, развития жизненного цикла и смены языковых знаков, т.е. в стационарном процессе функционирования языка). Эти законы распределения должны проявляться во временных сечениях процесса функционирования языка. Именно на таких сечениях могут быть проверены (путем сравнения со словарями или путем анализа текстов) предполагаемые законы распределения параметров, а также предсказаны путем компьютерного моделирования законы распределения характеристик, определяемых этими параметрами (частотного распределения полисемии, распределения уровней абстрактности значений знаков, времени жизни, возраста знаков и т.п.).

В стохастической модели развития во времени процессов функционирования больших ансамблей знаков на основе предложенной модели жизненного цикла отдельного языкового знака предполагается, что знаки возникают в языке в случайные моменты времени, образуя пуассоновский поток событий некоторой интенсивности. В этом случае интервал времени между появлением соседних по времени знаков языка имеет экспоненциальное распределение со средним значением, обратным интенсивности потока. Предполагается также, что постоянные времени $\tau_i^{(1)}$ и $\tau_i^{(2)}$ обратно пропорциональны G_i для всех знаков, а сами G_i образуют статистический ансамбль с экспоненциальным (точнее, с показательным, т.к. G_i могут принимать только целочисленные значения) законом распределения. Предполагается, что задержки τ_{0i} начала выхода значений знаков из употребления также подчиняются экспоненциальному закону распределения и не зависят статистически от значений G_i. Кроме того, в статистической модели предполагается, что моменты $t_{i,k}^{(1)}$ рождения новых значений и $t_{i,k}^{(2)}$ выхода из употребления этих значений также флуктуируют для каждого знака и значения, но так, что рождение значения более высокого уровня $k+1$ происходит не раньше рождения значения предыдущего, более низкого уровня k, то есть так, что $t_{i,k+1}^{(1)} \ge t_{i,k}^{(1)}$, и аналогично $t_{i,k+1}^{(2)} \ge t_{i,k}^{(2)}$, $\forall k$, $\forall i$. Эти флуктуации в модели описываются незави-

симыми равномерно распределенными случайными величинами с нулевыми средними значениями и полуширинами

$$t_{i,k+1}^{(1,2)} - t_{i,k}^{(1,2)} = \tau_i^{(1,2)}/(G_i - k)$$

плотностей распределения соответственно.

4 Статистическое моделирование эволюции языковых знаков

4.1 Параметры стохастической модели

Предложенная выше стохастическая диссипативная динамическая модель эволюции языковых знаков содержит 5 параметров:

- средний интервал времени $\langle\tau\rangle$ между соседними знаками в пуассоновском потоке знаков (слов), рождающихся в процессе эволюции языка (величина, обратная интенсивности потока);

- среднее значение $\langle G\rangle$ АСП знаков (параметр показательного распределения АСП);

- коэффициент $c^{(1)}$ обратно пропорциональной зависимости постоянной времени $\tau^{(1)} = c^{(1)}\langle G\rangle/G$ роста новых значений каждого знака от его нормированного АСП $G/\langle G\rangle$ (параметр, характеризующий скорость рождения новых значений знаков);

- коэффициент $c^{(2)}$ обратно пропорциональной зависимости постоянной времени $\tau^{(2)} = c^{(2)}\langle G\rangle/G$ выхода из употребления значений каждого знака от его нормированного АСП $G/\langle G\rangle$ (параметр, характеризующий скорость выхода значений знаков из употребления, $c^{(2)} >> c^{(1)} > 0$);

- среднее время $\langle\tau_0\rangle$ запаздывания начала выхода значений знака из употребления относительно момента появления знака в языке (параметр экспоненциального распределения запаздывания).

Значения этих параметров определяют конкретный вид результатов статистического моделирования эволюции языковых знаков. Однако математическая модель может быть полезной только тогда, когда она достаточно адекватно предсказывает реальные распределения знаков по числу значений (полисемии), по длительности жизни, по возрасту и т.д., т.е. когда результаты моделирования согласуются с реальными данными представительных словарей того или иного языка или с данными представительных корпусов текстов.

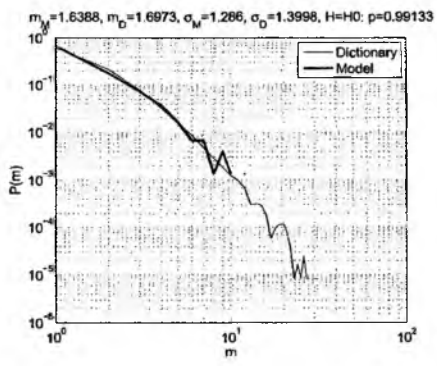

Рис. 1: Распределение полисемии по ССРЛЯ.

4.2 Идентификация модели и проверка ее адекватности по распределениям полисемии, полученным из представительных словарей

Для проверки адекватности модели использовалось 5 представительных словарей русского и английского языков (3 русского и 2 английского):

ССРЛЯ – "Словарь современного русского литературного языка" в 17-ти т.т. (1948-1965), (большой словарь);

МАС – "Словарь русского языка" под ред. А.П. Евгеньевой (1957-1961), (средний словарь);

СО – "Словарь русского языка" С.И. Ожегова в 4-х т.т. (1972, 9-е издание), (краткий словарь);

Shorter – "Shorter Oxford English Dictionary" (1962), (средний словарь);

Hornby – A.S. Hornby. "Oxford Advanced Learner's Dictionary of Current English" (1982), (краткий словарь).

На рис. 1 – 2 в качестве примера приведены в логарифмическом масштабе по обеим осям эмпирические (по словарям) и теоретические (по временным сечениям потока знаков, получаемых в модели) распределения актуальной полисемии знаков (слов) русского и английского языков при соответствующем подборе параметров модели (ее идентификации). Моделировалась динамика развития ансамбля 5000 знаков. Среднее значение интервалов между появлением знаков выбиралось равным $\langle\tau\rangle = \langle t_{i+1} - t_i\rangle = 1$ (принималось за единицу измерения времени). Идентификация по остальным четырем параметрам модели $\langle G\rangle$, $c^{(1)}$, $c^{(2)}$, $\langle\tau_0\rangle$ производилась методом стохастической аппроксимации [3]. При этом проверялась нулевая гипотеза H_0 об идентичности эмпирического и теоретического законов распределения полисемии

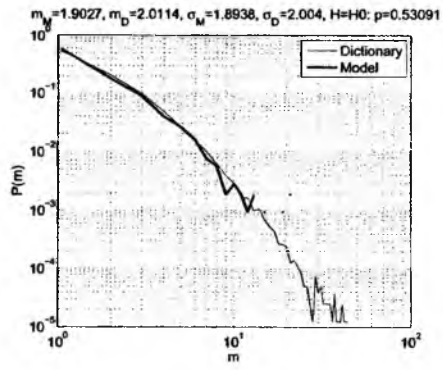

Рис. 2: Распределение полисемии по Shorter.

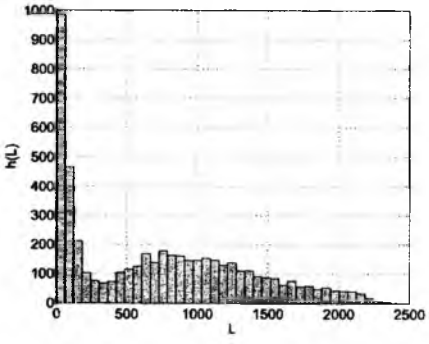

Рис. 3: Гистограмма распределения длительности жизни знаков.

против альтернативной гипотезы об их различии (по критерию Колмогорова-Смирнова). Во всех случаях уровень значимости критерия был $p >> 5\%$, так что нулевая гипотеза H_0 не отвергалась.

Как видно из приведенных рисунков, кривые распределений полисемии, полученные по предложенной диссипативной стохастической динамической модели эволюции языковых знаков, отлично согласуются с кривыми эмпирических распределений, полученными из представительных словарей русского и английского языков. По-видимому, можно ожидать согласия модельных распределений полисемии также и с данными представительных словарей других языков.

4.3 Распределение знаков по длительности жизни и возрасту

По временным сечениям (разрезам) статистического ансамбля (потока) моделируемых данных эволюции языковых знаков можно получить распределения не

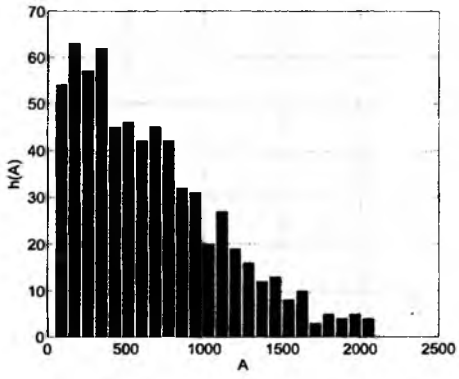

Рис. 4: Гистограмма распределения возраста знаков.

только полисемии знаков, но и длительности жизни знаков, возраста знаков и др.

На рис. 3– 4 в качестве примеров представлены одномоментные гистограммы распределений длительности жизни знаков и возраста знаков при значениях параметров модели, соответствующих большому словарю ССРЛЯ.

Насколько подобные модельные (теоретические) распределения длительностей жизни и возрастов знаков будут соответствовать эмпирическим распределениям, полученным из представительных словарей или больших корпусов текстов, покажут дальнейшие исследования.

Благодарности

Работа выполнялась при финансовой поддержке РФФИ, проект 10-01-00462а.

Список литературы

[1] А. С. Михайлов А. Ю. Лоскутов. *Основы теории сложных систем.* М.-Ижевск: Институт компьютерных исследований, 2007.

[2] А. А. Поликарпов. Системно-квантитативный подход в лингвистике. *Филологические школы и их роль в систематизации научных исследований*, Смоленск: Маджента:35–59, 2007.

[3] М. Вазан. *Стохастическая аппроксимация.* М.: Мир, 1972.

[4] М.А. Айзерман. *Классическая механика.* М.: Наука, Гл. ред. физ.-мат. лит., 1980.

[5] А.П. Маркеев. *Теоретическая механика: Учебник для университетов.* М.: ЧеРо, 1999.

XIV ЭМ'2010

Т р у д ы

XIV международной конференции по эвентологической математике и смежным вопросам

Под редакцией Олега Воробьёва

Отпечатано с готовых оригинал-макетов
Подписано в печать 06.12.2010 г. Формат 60 x 84/8
Бумага офсетная. Печать плоская.
Усл.-печ. л. 16.27 Уч.-изд. -л. 23.63
Тираж 150 экз. Заказ 496

Отпечатано в типографии ООО «Поликом»
Лицензия: серия НД № 06019 от 09.10.2001 г.
660093, г. Красноярск, ул. Академика Вавилова, 1, стр. 51, оф. 4-3
Тел.: (391) 285-85-17, тел/факс: (391) 276-80-10
E-mail: pkpolikom@mail.ru