ЭМПИРИЧЕСКОЕ ПРЕДСКАЗАНИЕ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1979 год Выпуск 79

УДК 8.74

ДИНАМИКА ИЗМЕНЕНИЯ ЭНТРОПИЙНЫХ ХАРАКТЕРИСТИК ТЕКСТА С РОСТОМ ЕГО ОБЪЕМА И ПОРЯДКА МОДЕЛИ

В.Д.Гусев, Т.Н.Титкова

В работах [1,2] авторами (совместно с Ю.Г.Косаревым) был предложен алгоритм получения распределений по частоте встречаемости l-грамм (связных подпоследовательностей из l символов) для произвольных символьных последовательностей значительной длины ($N \sim 10^6$ символов) и произвольных значений l ($l \le N$). Статистики, полученные для значений $l = 1,2,3,\ldots$, можно трактовать как оценки вероятностей появления в тексте всевозможных l-буквенных сочетаний (емкость соответствующего класса событий S_l равна n^l, где n - мощность алфавита). На основе этих статистик могут быть построены последовательно усложняющиеся модели текста в виде марковских цепей первого, второго и более высоких порядков.

При построении подобных моделей возникает вопрос о достоверности статистик, положенных в их основу. Для ответа на указанный вопрос в данной работе предлагается использовать в некоторых ситуациях энтропийный подход, суть которого состоит в том, чтобы привлекать в качестве эксперта самого человека с его, как правило, хорошим, но зачастую неформализованным знанием предметной области (в данном случае языка). Изложение будет иллюстрироваться на примере естественного языка.

1. Существующие методики оценки достоверности статистик. Очевидно, что ввиду ограниченности объема выборки (длины текста) отождествление частот с вероятностями возможно лишь до определенных значений $l = l^*(N)$. Заметим, что если текст не получен повторением одной и той же подпоследовательности, то с увеличением l число единичных (т.е. однократно встречающихся) l-грамм в тексте стремится к (N-l+1) -

полному числу 1-грамм в тексте. Это означает, что оценка условной энтропии текста (см.ниже) стремится к нулю, а оценка избыточности - к единице.

В традиционной вычислительной лингвистике [3,4] для оценки достоверности результатов, т.е. для вычисления относительной ошибки $\delta = |P-F|/P$, с которой по выборке объема N определяется частота F (оценка вероятности P) какой-либо лингвистической единицы (либо решается обратная задача - определяется N по заданной δ), используется выражение $\delta = Z_\alpha/\sqrt{NP}$, где Z_α - пороговая константа, фиксируемая выбором уровня значимости α (например, при $\alpha = 0.95$ константа $Z_\alpha \approx 2$). При этом предполагается, что текст можно представить в виде последовательности н е з а в и с и м ы х испытаний, и, как следствие, что частоты при больших N распределены по нормальному закону. Оба предположения не укладываются в рассматриваемую нами схему зависимых испытаний.

Другой возможный подход заключается в использовании оценок равномерного уклонения частот от вероятностей по конечному классу событий S_1 (см.[5]). Если потребовать выполнения неравенства

$$P\{\max_i |P(A_i^{(1)}) - P(A_i^{(1)})| > \varepsilon\} \le \eta, \qquad (1)$$

где $A_i^{(1)}$ - событие из класса S_1 ($1 \le i \le |S_1|$), то вытекающая отсюда связь между требуемым объемом выборки N, мощностью класса событий S_1 и параметрами ε и η имеет вид

$$N = \frac{\ln |S_1| - \ln \eta}{2\varepsilon^2}$$

или с учетом того, что $|S_1| = n^1$,

$$N = \frac{1 \cdot \ln n - \ln \eta}{2\varepsilon^2}. \qquad (2)$$

Однако оценки требуемого объема выборки, полученные с помощью (2), оказываются сильно завышенными. Действительно, любой осмысленный выбор параметра ε в (1) должен учитывать ожидаемые значения относительных частот $F(A_i^{(1)})$ событий класса S_1. Если потребовать, в частности, чтобы значения ε хотя бы на порядок уступали значениям частот, то, например, уже для случая биграмм (1 = 2) имеем: $F_{max}^{(2)} \approx 1{,}5 \cdot 10^{-2}$ (литературные данные и наши собственные эксперименты), $\varepsilon = 1{,}5 \cdot 10^{-3}$, $n = 32$, $\eta = 0.1$, откуда $N \approx 2 \cdot 10^6$. При $1 > 2$ значения частот быстро убывают (к примеру, $F_{max}^{(6)} \approx 0{,}1\, F_{max}^{(2)}$) и со-

ответственно в квадратичной пропорции растет требуемый объем выборки.

Заметим, что в классической лингвостатистике достаточным для набора статистики биграмм обычно считается текст длиной в несколько десятков тысяч символов ($N \sim 3 \cdot 10^4$). При этом средняя частота встречаемости биграмм в тексте примерно равняется 30. Близкими характеристиками обладают и частотные словари русского языка (средняя частота вхождения словоформы в выборку для словаря Л.Н.Засориной равняется, например, 27).

2. Э н т р о п и й н ы й п о д х о д. В данной работе предлагается оценивать параметр $1^*(N)$ косвенным образом путем сопоставления энтропии H_1 (или избыточности R_1), вычисленных двоя - ким образом: I) на основании полученных статистик и 2) используя результаты шенноновских экспериментов по угадыванию текста. Предполагается, что оценки, получаемые вторым способом, достаточно адекватно отражают истинные значения энтропии и избыточности, характеризующие язык в целом. Значения $1(N)$, при которых величины H_1, вычисленные по обеим методикам, начинают существенно отличаться, принимаются за оценки параметра $1^*(N)$.

К.Шеннон [6] определил условную энтропию H_1 порядка (1-1), т.е. энтропию, отвечающую опыту по выявлению 1-й буквы текста при наличии информации о предыдущих (1-1)-й буквах, в виде

$$H_1 = - \sum_{i,j} P(b_i,j) \log_2 Pb_i(j), \qquad (3)$$

где b_i - блок из (1-1)-й буквы ((1-1)-грамма); j - произвольная буква, следующая за b_i; $P(b_i,j)$ - вероятность встречаемости 1-граммы $b_i j$; $Pb_i(j) = P(b_i,j)/P(b_i)$ - условная вероятность следования буквы j за блоком b_i. Соответственно избыточность (1-1) - го порядка имеет вид

$$R_1 = 1 - \frac{H_1}{\log_2 n}. \qquad (4)$$

Энтропия и избыточность языка определяются как

$$H = \lim_{1 \to \infty} H_1; \quad R = \lim_{1 \to \infty} R_1.$$

Для вычислительных целей удобнее заменить (3) на эквивалентное выражение

$$H_1 = - \sum_{i,j} P(b_i,j) \log_2 P(b_i,j) + \sum_i P(b_i) \log_2 P(b_i) = H^{(1)} - H^{(1-1)}, \qquad (5)$$

где $H^{(1)}$ - энтропия 1-го порядка.

Подставляя в (5) оценки вероятностей

$$\hat{P}(b_i, j) = \frac{f(b_i, j)}{N-l+1}, \quad \hat{P}(b_i) = \frac{f(b_i)}{N-l+2},$$

где $f(b_i, j)$ и $f(b_i)$ - абсолютные частоты встречаемости всевозможных l- и $(l-1)$-грамм текста длины N, можно получить оценки для H_l. Вычисления строятся итеративно, начиная с $l = 2$.

Пользуясь выражением (5) и статистиками биграмм и триграмм английского языка, Шеннону удалось получить оценки условной энтропии для $l = 2$ и 3. Поскольку статистик более высокого порядка для английского языка к тому времени получено не было *), ввиду большой трудоемкости задачи, К.Шеннон предложил методику оценивания условной энтропии (и соответственно избыточности) языка, основанную на экспериментах по угадыванию человеком-носителем языка очередной буквы текста при условии, что предыдущие $(l-1)$ буквы ему известны. Подсчитывая для каждого символа текста число попыток, требовавшихся испытуемому для получения правильного ответа, К. Шеннон переходил от исходного текста к цифровому, основываясь на котором можно получить верхние ($\overline{H}_l$) и нижние ($\underline{H}_l$) оценки условной энтропии H_l исходного текста.

Следует подчеркнуть, что использование в качестве эксперта именно человека с его огромными, но лишь частично формализованными сведениями о статистике языка позволяет обойти трудности, связанные с отсутствием статистик l-грамм высокого порядка. С вычислительной точки зрения переход от исходного текста к цифровому эквивалентен переходу от l-мерной задачи к одномерной, поскольку верхние и нижние оценки для H_l вычисляются лишь на основании одномерных $(l = 1)$ статистик цифрового текста.

Эксперименты по оценке энтропии и избыточности были проделаны для многих языков с привлечением большого количества испытуемых [7] и в подавляющем большинстве случаев дали хорошо совпадающие результаты. Полученные оценки были подтверждены и при использовании других методик, в основу которых также было положено участие человека-эксперта (или коллектива экспертов). Это дает нам основание воспользоваться в своих экспериментах уже имеющимися верхними и нижними оценками условной энтропии при разных значениях l (см. [7]).

---

*) Насколько нам известно, таких данных не существует и в настоящее время.

3. О п и с а н и е э к с п е р и м е н т а. При помощи алгоритма [I] был получен полный спектр статистик ($l = I,2,3$ и т.д. вплоть до значения L, при котором в тексте уже не обнаруживалось ни одной повторяющейся L-граммы) для пяти текстов T_i с $N_i = i \cdot 10^5$ символов ($i = I,2,3,4,5$). Выборку составляли технические тексты по радиоэлектронике, причем $T_1 \subset T_2 \subset T_3 \subset T_4 \subset T_5$. Использовался алфавит из 35 символов (32 буквы, точка, запятая, пробел). В соответствии с (4) и (5) для каждого из текстов вычислялись значения R_l и H_l ($l = I,2,3,\ldots, L$). Результаты эксперимента совместно с верхними ($\overline{H}_l$) и нижними ($\underline{H}_l$) оценками условной энтропии [7] сведены в табл.I. Для сопоставления там же приведены энтропийные характеристики для двух текстов несколько отличной природы (T_6 - художественный текст, $N = I0^5$ символов, и T_7 - газетный текст на общественно-политическую тематику, $N = I0^5$ символов).

С энтропийными характеристиками тесно связаны такие параметры текста, как число единичных l-грамм E_l^1 и число различных кратных l-грамм $\sum_{k\neq 1} E_l^k$, где E_l^k - количество разновидностей l-грамм, каждая из которых встретилась в тексте ровно "k" раз. Все они приведены в табл. 2.

4. О б с у ж д е н и е р е з у л ь т а т о в э к с п е р и м е н т а.

4.I. При $l = 2$ полученные нами выборочные значения условной энтропии значительно превышают границу $\overline{H}_2$, полученную в экспериментах по угадыванию. Это, по-видимому, объясняется тем, что человек как эксперт еще неуверенно действует в ситуации, когда количество букв, предшествующих угадываемой и сообщаемых испытуемому, невелико (в данном случае одна). Пожалуй, лучшее, что он может предпринять в данной ситуации, это воспользоваться какой-либо имеющейся статистикой биграмм. В нашем случае выборка для получения соответствующей статистики была достаточно представительной.

4.2. При $l \sim 5$ и выше выборочные значения условной энтропии, приведенные в табл.I, несколько занижены (соответственно для избыточности завышены) по сравнению со значениями, которых следовало бы ожидать при совпадении условий данного эксперимента с теми, при которых были получены нижние и верхние оценки для H_l.

Первое отличие заключается в том, что эксперимент с увеличением N проводился на техническом тексте, а эксперименты по угадыванию - на художественном [7]. Поскольку технический текст яв-

Таблица I

Выборочные значения условной энтропии и избыточности в зависимости от длины текста и порядка модели (третий знак выписан с учетом округления)

l	$\underline{H}_1$	$\overline{H}_1$	Технический текст										Художественный текст		Общественно-политический текст	
			$\hat{H}_1(T_1)$	$\hat{R}_1(T_1)$	$\hat{H}_1(T_2)$	$\hat{R}_1(T_2)$	$\hat{H}_1(T_3)$	$\hat{R}_1(T_3)$	$\hat{H}_1(T_4)$	$\hat{R}_1(T_4)$	$\hat{H}_1(T_5)$	$\hat{R}_1(T_5)$	$\hat{H}_1(T_6)$	$\hat{R}_1(T_6)$	$\hat{H}_1(T_7)$	$\hat{R}_1(T_7)$
2	I,94	2,85	3,58	0,30	3,58	0,30	3,60	0,30	3,66	0,29	3,67	0,28	3,7I	0,27	3,58	0,30
3	2,II	3,I6	2,55	0,50	2,57	0,49	2,6I	0,49	2,75	0,46	2,76	0,46	3,0I	0,4I	2,68	0,48
4	I,55	2,45	I,66	0,67	I,76	0,66	I,77	0,65	I,92	0,62	I,95	0,6I	2,II	0,58	I,75	0,65
5	I,52	2,28	I,I3	0,78	I,22	0,76	I,29	0,75	I,38	0,73	I,43	0,72	I,3I	0,74	I,I5	0,77
6	I,56	2,54	0,79	0,84	0,92	0,82	0,97	0,8I	I,02	0,80	I,07	0,79	0,76	0,85	0,77	0,85
7	I,28	2,I2	0,57	0,89	0,68	0,87	0,74	0,85	0,76	0,85	0,80	0,84	0,43	0,9I	0,52	0,89
8	I,34	2,22	0,42	0,92	0,52	0,90	0,57	0,89	0,58	0,89	0,6I	0,88	0,25	0,95	0,37	0,92
9	I,30	2,05	0,32	0,94	0,45	0,9I	0,46	0,9I	0,44	0,9I	0,46	0,90	0,I6	0,97	0,28	0,94
I0	I,02	I,68	0,25	0,95	0,26	0,95	0,35	0,93	0,34	0,93	0,36	0,90	0,I0	0,98	0,2I	0,95
I00	0,58	I,02	0	I	0	I	0	I	0	I	0	I	0	I	0	I

Таблица 2

Число единичных (E_1^1) и различных кратных ($\sum_{k\neq 1} E_1^k$) 1-грамм в тексте в зависимости от его длины и порядка модели

1	Технический текст										Художественный текст		Общественно-политический текст	
	$E_1^1(T_1)$	$\Sigma E_1^k(T_1)$	$E_1^1(T_2)$	$\Sigma E_1^k(T_2)$	$E_1^1(T_3)$	$\Sigma E_1^k(T_3)$	$E_1^1(T_4)$	$\Sigma E_1^k(T_4)$	$E_1^1(T_5)$	$\Sigma E_1^k(T_5)$	$E_1^1(T_6)$	$\Sigma E_1^k(T_6)$	$E_1^1(T_7)$	$\Sigma_1^k(T_7)$
2	I3I	657	I28	7I0	I30	784	II4	890	I02	9I8	82	738	I04	680
3	I478	3572	I642	4287	2082	4960	2661	6393	2638	6792	I627	52I2	I509	3895
4	63I3	8437	I24I9	942I	9395	I4503	I4239	I9265	I5055	2I465	I0776	I2886	70I7	92I3
5	I573I	I2I26	20776	I9760	26287	25628	39769	34063	43749	39935	29II2	I6076	I7823	I29I9
6	27437	I3388	39922	24462	5I778	33390	7699I	43974	87726	5335I	48545	I4476	3II82	I356I
7	385I3	I3343	60428	26222	80595	37026	II7642	48I62	I37343	59769	63799	II509	43314	I2578
8	48503	I2543	80045	26007	I08957	37805	I56603	4860I	I85876	6I389	7465I	8659	532II	II263
9	57064	II487	I08545	I9326	I35I27	36805	I9I774	46867	230381	59907	82093	64I2	6I246	9926
I0	64423	I0320	II3392	2289I	I58566	34862	222624	43996	269607	56769	873I4	4682	678I5	87I5
I00	10^5	0	$2\cdot10^5$	0	$3\cdot10^5$	0	$4\cdot10^5$	0	$5\cdot10^5$	0	10^5	0	10^5	0

ляется более избыточным, чем художественный, для него границы $\underline{H}_1$ и $\overline{H}_1$ должны быть несколько снижены. Величину расхождения можно оценить при сопоставлении текстов T_1 и T_6.

Другое отличие заключается в разной мощности алфавита ($n = 35$ в рассматриваемом эксперименте и $n = 32$ в эксперименте по угадыванию). Три дополнительных символа (пробел, запятая, точка), используемых нами, при больших l легко предсказуемы, что уменьшает неопределенность предсказания по сравнению со случаем их отсутствия. Соответствующий вопрос рассматривался в [8], где получено соотношение, связывающее $H = H_\infty$ для текстов с пробелами и без них. Применительно к нашему случаю данное соотношение будет иметь вид: $H(n = 35) = (1-P) \cdot H\ (n = 32)$, где P - суммарная вероятность встречаемости трех дополнительных символов. Поскольку основной вклад в сумму вносит вероятность пробела ($\hat{P}(\sqcup) = 0{,}095$ для технического текста, примерно 0,12 - для общественно-политического и 0,13 - для художественного), поправка к $\hat{H}_1$ (в сторону ее увеличения) при больших l должна составлять порядка одной десятой от значения $\hat{H}_1$.

С учетом указанной поправки выборочные значения энтропии для самого длинного текста T_5 ($N = 5 \cdot 10^5$ символов) вкладываются в "шенноновский коридор" лишь для диапазона значений $l = 3 - 5$, т.е. $l^*(N = 5 \cdot 10^5) = 5$.

4.3. Скорость сходимости частот к вероятностям довольно медленная: пятикратному увеличению объема текста при переходе от T_1 к T_5 соответствует увеличение параметра $l^*(N)$ всего лишь на единицу ($l^*(N = 10^5) = 4$, $l^*(N = 5 \cdot 10^5) = 5$).

4.4. Интересный эффект выявляется при сопоставлении значений избыточности для художественного и общественно-политического (или технического) текстов. Естественно ожидать для художественного текста меньших значений избыточности, чем для общественно-политического и технического текстов, представляющих отдельные подъязыки языка в целом. Для значений $l \leq 5$, т.е. в диапазоне значений, когда мы можем считать наши статистики достоверными в указанном выше смысле, $R_l^{\text{худ.}} < R_l^{\text{общ.-пол.}}$, что согласуется с нашими интуитивными представлениями. При $l \geq 6$ картина меняется на противоположную. Это не означает, однако, что художественный текст стал более избыточным, а свидетельствует о том, что для менее избыточного текста эффект недостаточности объема выборки с ростом l сказывается сильнее, чем для более избыточного.

Таким образом, если положить в основу посылку о том, что одна кривая избыточности, рассматриваемая как функция от l при фик-

сированном N, мажорирует другую при значениях N, обеспечивающих достоверность статистик, то пересечение этих кривых при некотором значении l будет эквивалентно невыполнению предпосылки о достаточности объема выборки для данного и всех последующих значений l. Точка пересечения дает дополнительную возможность для оценивания параметра $l^*(N)$.

4.5. Число единичных l-грамм в тексте с увеличением l (при фиксированном N) монотонно возрастает. Зависимость $E_1^1(l)$ в диапазоне значений $4 \le l \le 10$ хорошо аппроксимируется линейной функцией для всех значений N. Скорость нарастания выше для меньших значений N (к примеру, $E_{10}^1/N = 0{,}64$ для T_1 и $0{,}54$ для T_5).

4.6. Количество различных кратных l-грамм ($\sum_{k \neq 1} E_l^k$) с увеличением l вначале возрастает, затем падает до нуля (при $l \sim 50 - 100$). Максимум приходится на значения l в диапазоне от 5 до 8. С увеличением N максимум сдвигается в сторону больших l.

5. З а м е ч а н и я.

5.1. Было бы неверным считать, что при $l > l^*(N)$ значения всех частот становятся статистически недостоверными. Интуитивно ясно, что для достижения одной и той же относительной ошибки $\delta = |P-F|/P$ при оценке вероятности различных лингвистических событий объем выборки должен быть тем больше, чем меньше P. Если N фиксировано, то для части наиболее вероятных событий данный объем выборки может оказаться достаточным, в то время как для маловероятных событий потребуется больший объем.

Расстановка "доверительных границ" внутри каждого из упорядоченных по убыванию частотных спектров l-грамм при $l > l^*(N)$ не является целью данной работы, однако можно указать некоторые соображения, которые могли бы быть положены в основу подобной процедуры.

Кажется естественным считать объем выборки N_1 достаточным для наиболее высокочастотной части словаря, ограниченной первыми R_1 членами, если при увеличении объема выборки до N_2, где $N_2 - N_1$ одного порядка с N_1, получается словарь, первые R_1 членов которого в некотором смысле близки к выделенным R_1 элементам первого словаря. Укажем три меры близости, подходящие для сравнения частотных словарей l-грамм.

Наиболее грубой мерой, учитывающей лишь близость словарей по составу входящих в них l-грамм, является

$$\beta_1^{(1)} = \frac{1}{R_1} \sum_{i,j=1}^{R_1} \delta_{ij}, \quad l = 1,2,3,\ldots, \tag{6}$$

где индекс i пробегает по элементам первого словаря, индекс j - по элементам второго, а

$$\delta_{ij} = \begin{cases} 1, \text{ если } i\text{-я } l\text{-грамма первого словаря совпадает с } j\text{-й} \\ \quad l\text{-граммой второго словаря;} \\ 0 - \text{в противном случае.} \end{cases}$$

Более тонкой (но соответственно менее устойчивой) является мера, учитывающая порядок l-грамм в словаре:

$$\rho_1^{(2)} = \sum_{i=1}^{R_1} |\, i - j_i \,|, \quad l = 1,2,3,\ldots, \tag{7}$$

где i - ранг l-грамм в первом словаре, j_i - ранг той же l-граммы во втором словаре. Данная мера является естественной (на случай l-граммных словарей) модификацией аналогичной меры для слов [4]. Мера (7) является аналогом расстояния ($\rho_1^{(2)} = 0$, если словари совпадают). Переход к мере близости может быть осуществлен заменой $\rho_1^{(2)}$ на $\beta_1^{(2)} = 1 - \rho_1^{(2)}/\rho_{1_{max}}^{(2)}$.

Еще более тонкая (и соответственно еще менее устойчивая) мера получается при использовании значений частот:

$$\rho_1^{(3)} = \frac{1}{R_1} \cdot \sum_{i=1}^{R_1} \left| \frac{f_i}{N_1} - \frac{f_{j_i}}{N_2} \right|, \quad l = 1,2,3,\ldots, \tag{8}$$

где f_i - абсолютная частота i-й по порядку l-граммы первого словаря, f_{j_i} - абсолютная частота той же l-граммы во втором словаре. Переход от (8) к соответствующей мере близости $\beta_1^{(3)}$ может быть осуществлен аналогично предыдущему случаю.

Для иллюстрации вышеизложенного в табл.3 приведены начала частотных рядов для текстов T_1, T_2, T_3 ($l = 3 < l^*(N)$). Аналогичные данные для $l = 10 > l^*(N)$ приведены в табл. 4. Привязка по рангам сделана к тексту T_1, т.е. за основу берутся 7 первых l-грамм частотного словаря этого текста за исключением тех из них, которые входят в состав одного слова (пропуски в табл. 4).

Анализ табл.3 показывает, что при $l < l^*(N)$ частотные спектры всех трех текстов близки друг другу по любому из трех критериев. При $l > l^*(N)$ (табл.4) частотные спектры сильнее отличаются друг от друга, однако явно заметен эффект стабилизации спектров с увеличением N. Легко видеть, что близость между частотными спект-

Таблица 3

Частотные спектры триграмм для текстов T_1, T_2, T_4

Триграмма	T_1		T_2		T_4	
	f	r	f	r	f	r
ЕНИ	822	I	I627	I	28I3	I
␣ПО	60I	2	I377	2	24I9	2
ОСТ	483	3	II44	3	2234	3
ПОЛ	478	4	946	4	I467	6
ТЬ␣	447	5	90I	5	I424	7
␣ПР	4II	6	832	6	I644	4
НИЯ	378	7	723	9	I276	I3

Таблица 4

Частотные спектры I0-грамм для текстов T_1, T_2, T_4

I0-граммы	T_1		T_2		T_4	
	f	r	f	r	f	r
␣ПОТЕНЦИАЛ	86	I	I77	3	258	3
ПОВЕРХНОСТ	77	2	267	I	359	I
СООТВЕТСТВ	64	4	I24	5	2I2	4
СООТНОШЕНИ	57	5	78	2I	II3	24
␣ВЫРАЖЕНИЕ	53	6	70	28	I04	37
ОМ␣СЛУЧАЕ,	50	7	69	32	88	49
РОСТРАНСТВ	43	I3	74	24	III	26

рами текстов T_2 и T_4 выше, чем между спектрами T_1 и T_2, хотя объем выборки в обоих случаях (т.е. при переходе от T_1 к T_2 и от T_2 к T_4) увеличивался в равной пропорции ($N_2/N_1 = N_4/N_2 = 2$). Все три текста достаточно близки друг другу по 1-граммному составу (мера (6)). Большой разброс наблюдается по рангам (мера (7)) и еще бо́льший - по значениям частот (мера (8)).

5.2. Анализ текста, использующий 1-граммный подход, не теряет своей значимости и вне рамок статистической модели. Он может быть применен к тексту любой длины и произвольной природы и позволяет зачастую выявить интересные локальные и интегральные закономерности, характеризующие внутреннюю структуру данного конкретного текста. В частности, 1-граммный анализ длинного текста является основой для сжатия его без потери информации, а 1-граммный анализ короткого текста может быть положен в основу формирования различных мер близости (аналогичных, например, (6) - (8)) для двух символьных последовательностей.

Литература

1. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. О задаче поиска повторяющихся отрезков текста. -В кн.: Вычислительные системы. Вып. 62. Ассоциативное кодирование. Новосибирск, 1975, с.49-71.

2. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. Отыскание статистических закономерностей текстов методом ассоциативного кодирования. -В кн: Вычислительные системы. Вып. 62. Ассоциативное кодирование. Новосибирск, 1975, с.72-89.

3. О точных методах исследования языка. /Ахманова О.С., Мельчук И.А., Падучева Е.В., Фрумкина Р.М. М.: Изд. МГУ, 1961.

4. КАЛИНИНА Е.А. Изучение лексико-статистических закономерностей на основе вероятностной модели. -В кн.: Статистика речи. - Л., Наука, 1968, с.64-107.

5. ВАПНИК В.Н., ЧЕРВОНЕНКИС А.Я. Теория распознавания образов. -М.: Наука, 1974.

6. ШЕННОН К. Предсказание и энтропия печатного английского текста. -В кн.: Работы по теории информации и кибернетике. М., 1963, с. 669-686.

7. ПИОТРОВСКИЙ Р.Г. Информационные измерения языка. -Л.: Наука, 1968.

8. ЯГЛОМ А.М., ЯГЛОМ И.М. Вероятность и информация. -М.: Физматгиз, 1960.

Поступила в ред.-изд.отд.
19 июня 1979 года