Дискретная математика

том 26 выпуск 2 * 2014

УДК 519.713.22 DOI 10.4213/dm1281

О мощности биграммных языков

© 2014 г. А. А. Петюшко*

Рассматриваются формальные языки, заданные матрицей кратностей биграмм и изученные на качественном уровне в опубликованной ранее статье А. А. Петюшко „О биграммных языках“. Исследуется вопрос зависимости мощности языка от исходной матрицы кратностей биграмм. Находятся асимптотические оценки для мощности языков и отношения количества матриц кратностей биграмм с различными свойствами.

Ключевые слова: матрица кратностей биграмм, биграммные языки, частотные языки, регулярность языков, эйлеровы циклы, ориентированные графы, мощность языка.

1. Введение

Еще в начале 20 века выдающимся русским ученым А. А. Марковым был создан аппарат цепей, впоследствии названных цепями Маркова, и опробован [1] на вычислении переходных вероятностей между соседними буквами (биграммами) в поэме А. С. Пушкина „Евгений Онегин“. В дальнейшем этот аппарат получил широкое применение для распознавания и статистического моделирования естественных языков [2]. Тем не менее, в детерминированном случае, за редким исключением прикладных задач (например, для подсчета ДНК-последовательностей [3]), биграммы для исследования формальных языков практически не применялись. В данной статье автор изучает языки, состоящие из слов с фиксированными частотами пар соседних букв, в частности, их мощностные характеристики.

Введем основные определения, которые нам понадобятся в основной части статьи. За более подробным изложением тем, касающихся вводимых понятий, можно обратиться к работе [4].

Пусть A ($|A| < \infty$) — конечный алфавит. A^* — множество всех слов, включая пустое, A^+ — множество всех непустых слов в данном алфавите.

Определение 1. Биграммой в алфавите A называется двухбуквенное слово $ab \in A^*, a, b \in A$ (порядок вхождения букв в биграмму имеет значение, т.е. биграмма ab не равна биграмме ba при $a \neq b$).

Определение 2. Обозначим через $\theta_\beta(\alpha)$, где $\beta \in A^+, \alpha \in A^*$, отображение $A^+ \times A^* \to N \cup \{0\}$, сопоставляющее слову α число подслов β в слове α, т.е. количество различных разложений слова α в виде $\alpha = \alpha'\beta\alpha''$ (α' и α'' могут быть пустыми). При длине слова α, меньшей длины слова β, значение $\theta_\beta(\alpha)$ положим равным 0. Само же значение $\theta_\beta(\alpha)$ при данных β и α назовем кратностью β в слове α.

*Место работы: МГУ им. М. В. Ломоносова, e-mail: petsan@newmail.ru

С учетом введенных определений по каждому слову $\alpha \in A^*$ можно построить квадратную матрицу кратностей биграмм $\Theta(\alpha) = (\theta_{a_i a_j}(\alpha))_{i,j=1}^{|A|}$ размера $|A|\times|A|$ при условии, что все буквы алфавита $A = \{a_1, a_2, ..., a_{|A|}\}$ пронумерованы и нумерация зафиксирована.

Обозначим через Ξ множество квадратных матриц размера $|A| \times |A|$, каждый элемент которых является целым неотрицательным числом. Таким образом, для каждого $\alpha \in A^*$ имеем $\Theta(\alpha) \in \Xi$. Здесь и далее через $\Theta(\alpha)$ будем обозначать матрицу кратностей биграмм, построенную по конкретному слову α, а через Θ — просто некоторую матрицу из Ξ, при этом будем считать, что на месте (i, j) матрицы Θ стоит значение $\theta_{a_i a_j}$ (для произвольной матрицы из Ξ мы опустили зависимость от α как для самой матрицы кратностей биграмм, так и для отдельных ее элементов).

Определение 3. Назовем языком $L(\Theta)$, порожденным матрицей $\Theta \in \Xi$, множество всех слов, имеющих одну и ту же матрицу кратностей биграмм Θ, т.е. $L(\Theta) = \{\beta \in A^* | \Theta(\beta) = \Theta\}$.

Пример 1. Пусть $A = \{0, 1\}$, $\alpha = 01011100$.
Тогда матрица кратностей биграмм $\Theta(\alpha) = \begin{pmatrix} \theta_{00}(\alpha) & \theta_{01}(\alpha) \\ \theta_{10}(\alpha) & \theta_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.

Построим по матрице $\Theta(\alpha)$ (или по произвольной матрице $\Theta \in \Xi$) ориентированный граф $G_{\Theta(\alpha)}$ на плоскости. Вершинами у этого графа будут все буквы из алфавита A, при этом ребра будут соответствовать биграммам с учетом их кратностей, т.е. кратность $\theta_{ab}(\alpha)$ будет порождать $\theta_{ab}(\alpha)$ ориентированных ребер $a \to b$. Аналогично, кратность $\theta_{cc}(\alpha)$ будет порождать $\theta_{cc}(\alpha)$ петель $c \to c$.

Пример 2. $A = \{0, 1\}$, $\alpha = 01011100$.
$\Theta(\alpha) = \begin{pmatrix} \theta_{00}(\alpha) & \theta_{01}(\alpha) \\ \theta_{10}(\alpha) & \theta_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф $G_{\Theta(\alpha)}$ по $\Theta(\alpha)$ - см. рис. 1.

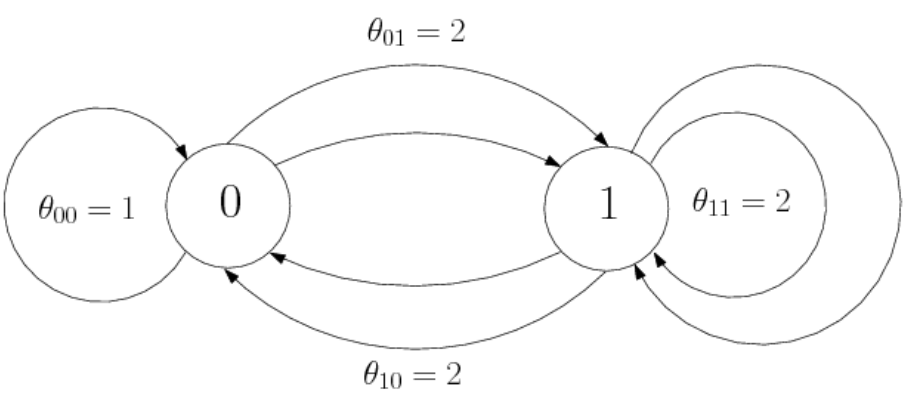

Рис. 1. Граф $G_{\Theta(\alpha)}$, построенный по набору $\Theta(\alpha)$

Напомним несколько широко известных понятий, касающихся эйлеровых путей.

Определение 4. Путем в ориентированном графе называется такая последовательность попарно различных ребер (набор параллельных ребер будем считать состоящим из различных ребер), что конец предыдущего ребра совпадает с началом следующего.

Определение 5. Циклом в ориентированном графе называется такой путь, что начало первого ребра в этом пути совпадает с концом последнего.

Определение 6. Эйлеровым путем в ориентированном графе называется такой путь, который содержит все ребра этого графа.

Определение 7. Эйлеровым циклом в ориентированном графе называется такой цикл, который содержит все ребра этого графа.

Определение 8. Полуэйлеров граф — граф, содержащий эйлеров путь, который не является эйлеровым циклом.

Определение 9. Эйлеров граф — граф, содержащий эйлеров цикл.

Замечание 1. На самом деле в каноническом определении полуэйлерова графа не говорится о том, что эйлеров путь не должен являться эйлеровым циклом. Но, следуя такому определению, несложно заметить, что любой эйлеров граф является также и полуэйлеровым, поэтому каждый раз для разграничения данных понятий пришлось бы дополнять фразой „полуэйлеров граф, не являющийся эйлеровым“.

В [5] доказаны следующие важные теоремы, позволяющие достаточно просто проверять ориентированные графы на наличие эйлеровых путей и циклов:

Теорема 1. *Ориентированный граф является эйлеровым тогда и только тогда, когда выполняются следующие условия*:

1) *все вершины, инцидентные ребрам, лежат в одной компоненте связности соответствующего неориентированного графа*;

2) *у всех вершин количество входящих ребер равно количеству исходящих ребер*.

Теорема 2. *Ориентированный граф является полуэйлеровым тогда и только тогда, когда выполняются следующие условия*:

1) *все вершины, инцидентные ребрам, лежат в одной компоненте связности соответствующего неориентированного графа*;

2) *у всех вершин, кроме двух, количество входящих ребер равно количеству исходящих ребер. У оставшихся двух вершин разность количества входящих ребер и количества исходящих ребер равна* +1 *и* -1 *соответственно*.

В дальнейшем, там, где будут упоминаться понятия эйлеровых и полуэйлеровых графов, будем иметь в виду, что установить факт, является граф эйлеровым или полуэйлеровым, можно по двум вышеприведенным теоремам.

Более интересен случай, когда мы рассматриваем матрицу кратностей биграмм не как абсолютное ограничение, а как задание относительных значений (пропорций) биграмм, т.е. случай языка, в котором отношения $\theta_{ab}(\alpha)/\theta_{cd}(\alpha)$ зависят только от букв $a, b, c, d \in A$, $\theta_{cd}(\alpha) > 0$, но не зависят от слова α из этого языка. Определим такой язык.

Определение 10. Назовем частотным языком на биграммах с кратностями, заданным матрицей кратностей биграмм $\Theta \in \Xi$, язык

$$F_\Theta = \bigcup_{k=1}^{\infty} L(k\Theta),$$

т.е. язык, состоящий из всех таких слов $\beta \in A^*$, что набор кратностей этих слов $\Theta(\beta)$ кратен набору Θ, а именно, $F_\Theta = \{\beta \in A^* | \Theta(\beta) = k\Theta, k \in N\}$, где умножение k на Θ понимается как умножение скаляра на матрицу.

2. Мощность языка $L(\Theta)$ и асимптотические оценки для языков $F(\Theta)$

Рассмотрим вопрос о том, сколько существует слов с данным набором Θ. Для начала рассмотрим случай двухбуквенного алфавита.

Теорема 3. *Для алфавита $A = \{0,1\}$ и матрицы $\Theta \in \Xi$, задающей эйлеров или полуэйлеров граф G_Θ, число слов N_Θ с заданной матрицей кратностей биграмм Θ:*

1) *при $\theta_{01} > \theta_{10}$ $N_\Theta = C^{\theta_{11}}_{\theta_{11}+\theta_{10}} C^{\theta_{00}}_{\theta_{00}+\theta_{10}}$;*
2) *при $\theta_{01} < \theta_{10}$ $N_\Theta = C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}}$;*
3) *при $\theta_{01} = \theta_{10}$ $N_\Theta = C^{\theta_{00}}_{\theta_{00}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} \left(\frac{\theta_{01}}{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}}\right)$; (здесь под C^k_n понимается число сочетаний из n по k, то есть $C^k_n = \frac{n!}{k!(n-k)!}$).*

Доказательство. Поскольку далее будет доказано более общее утверждение, здесь приведем лишь основную идею. Доказательство заключается в комбинаторном подсчете количества способов разместить однородные фрагменты из 0 и 1 по местам, где есть хотя бы одна такая буква. Также в случае 1) или 2) мы однозначно можем определить первую и последнюю буквы любого слова из $L(\Theta)$ (причем эти буквы — разные), в то время как в случае 3) множество искомых слов из $L(\Theta)$ распадается на два подмножества: одно со словами, начинающимися и заканчивающимися на 0, другое — на 1.

Пример 3. $A = \{0,1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Так как $\theta_{01} = \theta_{10}$, то искомое число слов с данным набором по теореме выше равно $N_\Theta = C^{\theta_{00}}_{\theta_{00}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} \left(\frac{\theta_{01}}{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}}\right) = C^1_3 C^2_4 \left(\frac{2}{3} + \frac{1}{2}\right) = 12 + 9 = 21$. Это действительно так, поскольку с данной матрицей кратностей Θ существует ровно 21 слово: 11100101, 11001101, 11001011, 10011101, 10011011, 10010111, 11101001, 11011001, 11010011, 10111001, 10110011, 10100111, 00111010, 00110110, 00101110, 01110010, 01100110, 01001110, 01110100, 01101100, 01011100.

В случае произвольного алфавита A нам понадобятся следующие определение и сопутствующая лемма.

Определение 11. Матрицей Кирхгофа $ML(\Theta)$ [6], построенной по матрице $\Theta \in \Xi$, называется квадратная матрица размером $|A| \times |A|$, в которой на месте (i,j) стоит элемент

$$l_{ij} = \begin{cases} -\theta_{a_i a_j}, & i \neq j; \\ \sum_{a_j \neq a_i} \theta_{a_i a_j}, & i = j. \end{cases}$$

Замечание 2. Матрица Кирхгофа для любой матрицы Θ имеет нулевой определитель ($\det ML(\Theta) = 0$), поскольку, очевидно, сумма всех столбцов $ML(\Theta)$ есть нулевой столбец.

Лемма 1. *Если матрица $\Theta \in \Xi$ такова, что соответствующий ориентированный граф G_Θ является эйлеровым, то все главные миноры $D^{(i,i)}$, полученные вычеркиванием из $ML(\Theta)$ i-й строки и i-го столбца, одинаковы.*

Доказательство. Пусть $|A| = n$. Без ограничения общности докажем, что $D^{(1,1)} = D^{(2,2)}$. Запишем

$$D^{(1,1)} = \begin{vmatrix} \sum_{a_j \neq a_2} \theta_{a_2 a_j} & -\theta_{a_2 a_3} & \dots & -\theta_{a_2 a_n} \\ -\theta_{a_3 a_2} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \dots & -\theta_{a_3 a_n} \\ \dots & \dots & \dots & \dots \\ -\theta_{a_n a_2} & -\theta_{a_n a_3} & \dots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix}.$$

Прибавим к первой строке все остальные строки (определитель при этом не изменится). Тогда на позиции $s-1$ ($s > 2$) в первой строке будет стоять $\sum_{a_j \neq a_s} \theta_{a_s a_j} - (\sum_{a_t \neq a_s} \theta_{a_t a_s} - \theta_{a_1 a_s})$. По Лемме 1 из [4] с учетом того, в эйлеровом цикле первая буква совпадает с последней, получаем $\sum_{a_j \neq a_s} \theta_{a_s a_j} = \sum_{a_t \neq a_s} \theta_{a_t a_s}$ при любом s. Значит,

$$D^{(1,1)} = \begin{vmatrix} \theta_{a_1 a_2} & \theta_{a_1 a_3} & \dots & \theta_{a_1 a_n} \\ -\theta_{a_3 a_2} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \dots & -\theta_{a_3 a_n} \\ \dots & \dots & \dots & \dots \\ -\theta_{a_n a_2} & -\theta_{a_n a_3} & \dots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix}.$$

Теперь прибавим к первому столбцу все остальные столбцы (определитель при этом не изменится). Тогда на позиции $s-1$ ($s > 3$) в первом столбце будет стоять $\sum_{a_j \neq a_s} \theta_{a_s a_j} - (\sum_{a_t \neq a_s} \theta_{a_s a_t} - \theta_{a_s a_1})$. Получим

$$D^{(1,1)} = \begin{vmatrix} \sum_{a_j \neq a_1} \theta_{a_1 a_j} & \theta_{a_1 a_3} & \dots & \theta_{a_1 a_n} \\ \theta_{a_3 a_1} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \dots & -\theta_{a_3 a_n} \\ \dots & \dots & \dots & \dots \\ \theta_{a_n a_1} & -\theta_{a_n a_3} & \dots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix}.$$

Умножим первую строку и первый столбец на (-1); определитель опять не изменится. Получим

$$D^{(1,1)} = \begin{vmatrix} \sum_{a_j \neq a_1} \theta_{a_1 a_j} & -\theta_{a_1 a_3} & \dots & -\theta_{a_1 a_n} \\ -\theta_{a_3 a_1} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \dots & -\theta_{a_3 a_n} \\ \dots & \dots & \dots & \dots \\ -\theta_{a_n a_1} & -\theta_{a_n a_3} & \dots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix} = D^{(2,2)}.$$

Таким образом, в случае эйлерова графа G_Θ можно рассматривать величину D, которая равна любому из миноров $D^{(i,i)}$ для любого $i = 1, \dots, |A|$.

Пусть $a \in A$, $\theta_a(\alpha)$ — количество (кратность) букв a (униграмм) в слове α, а $\Delta(\alpha) = (\theta_{a_1}(\alpha), \theta_{a_2}(\alpha), ..., \theta_{a_{|A|}}(\alpha))$ — вектор кратностей униграмм. Пусть Ξ_Δ - пространство векторов размера $|A|$, состоящих из неотрицательных целых чисел. Очевидно, что для любого слова $\alpha \in A^*$ $\Delta(\alpha) \in \Xi_\Delta$. Здесь и далее будем через Δ обозначать произвольный вектор из Ξ_Δ, при этом будем считать, что на месте i вектора Δ стоит значение θ_{a_i}.

Теорема 4. *Число слов в алфавите A с заданными вектором кратностей униграмм $\Delta \in \Xi_\Delta$ с положительными элементами и матрицей кратностей биграмм $\Theta \in \Xi$ есть*

$$N_{\Delta,\Theta} = \frac{\prod_{a_i \in A} (\theta_{a_i} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} \det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j})_{i,j=1}^{|A|},$$

где δ_{ij} — символ Кронекера [3].

В нашем случае нет данных о кратностях униграмм в дополнение к кратностям биграмм Θ. Поэтому предыдущая теорема нуждается в усилении.

Теорема 5. *Пусть задана матрица кратностей биграмм $\Theta \in \Xi$, которой соответствует эйлеров или полуэйлеров граф G_Θ, причем для каждого i существует такое $j \neq i$, что $\theta_{a_i a_j} > 0$ или $\theta_{a_j a_i} > 0$. Тогда*:

1) *если существует такое i', что $\sum_{a_i \in A} \theta_{a_i a_{i'}} > \sum_{a_i \in A} \theta_{a_{i'} a_i}$, то*

$$N_\Theta = \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1 + \delta_{i'i})!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')},$$

где $\delta_{i'i}$ — символ Кронекера;

2) *если для любых $i, j = 1, \ldots, n$ $\sum_{a_i \in A} \theta_{a_i a_j} = \sum_{a_i \in A} \theta_{a_j a_i}$, то*

$$N_\Theta = \left(\sum_{a_i, a_j \in A} \theta_{a_i a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D.$$

Доказательство. Точно найти вектор кратностей униграмм по матрице кратностей биграмм мы не сможем. Однако, поскольку нас интересует такая интегральная характеристика, как мощность множества, то это и не требуется.

Прежде всего отметим, что согласно Лемме 2 из [4] существует хотя бы одно слово α с матрицей кратностей биграмм Θ.

Далее заметим, что условие существования для каждого i такого $j \neq i$, что $\theta_{a_i a_j} > 0$ или $\theta_{a_j a_i} > 0$, означает, что каждая буква из алфавита A встретится хотя бы раз в написании слова, соответствующего матрице кратностей биграмм Θ, т.е. $\theta_{a_i} > 0$ для любого i, значит, выполняется условие Теоремы 4.

1) Пусть существует такое i', что $\sum_{a_i \in A} \theta_{a_i a_{i'}} > \sum_{a_i \in A} \theta_{a_{i'} a_i}$. Это означает, что граф G_Θ является полуэйлеровым и для любого такого слова $\alpha \in A^*$, что $\Theta(\alpha) = \Theta$, буква $a_{i'}$ будет являться последней буквой слова α (согласно Лемме 1 из [4]).

Более того, теперь можно вычислить однозначно кратность любой униграммы θ_{a_i}. Если $a_i \neq a_{i'}$, то $\theta_{a_i} = \sum_{a_j \in A} \theta_{a_i a_j}$. При этом $\theta_{a_{i'}} = \sum_{a_j \in A} \theta_{a_{i'} a_j} + 1$, поскольку кратность последней буквы на единицу больше количества исходящих ребер в соответствующем полуэйлеровом графе G_Θ.

Теперь преобразуем детерминант из формулировки Теоремы 4. Подставляя выведенные выше выражения для униграмм, получим, что

$$\det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j}) =$$

$$= \begin{vmatrix} \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \sum_{a_j \neq a_{i'-1}} \theta_{a_{i'-1} a_j} & -\theta_{a_{i'-1} a_{i'}} & -\theta_{a_{i'-1} a_{i'+1}} & \ldots \\ \ldots & -\theta_{a_{i'} a_{i'-1}} & \sum_{a_j \neq a_{i'}} \theta_{a_{i'} a_j} + 1 & -\theta_{a_{i'} a_{i'+1}} & \ldots \\ \ldots & -\theta_{a_{i'+1} a_{i'-1}} & -\theta_{a_{i'+1} a_{i'}} & \sum_{a_j \neq a_{i'+1}} \theta_{a_{i'+1} a_j} & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{vmatrix}.$$

Прибавим к i'-му столбцу сумму остальных столбцов (при этом определитель не изменится). Получим:

$$\det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j}) = \begin{vmatrix} \dots & \dots & 0 & \dots & \dots \\ \dots & \sum_{a_j \neq a_{i'-1}} \theta_{a_{i'-1} a_j} & 0 & -\theta_{a_{i'-1} a_{i'+1}} & \dots \\ \dots & -\theta_{a_{i'} a_{i'-1}} & 1 & -\theta_{a_{i'} a_{i'+1}} & \dots \\ \dots & -\theta_{a_{i'+1} a_{i'-1}} & 0 & \sum_{a_j \neq a_{i'+1}} \theta_{a_{i'+1} a_j} & \dots \\ \dots & \dots & 0 & \dots & \dots \end{vmatrix}.$$

Разложим определитель по элементам i'-го столбца. Так как определитель матрицы равен сумме произведений элементов столбца на их алгебраические дополнения, а единственный ненулевой элемент в столбце находится в строке i' и равен 1, то

$$\det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j}) =$$

$$= 1 * (-1)^{i'+i'} \begin{vmatrix} \dots & \dots & \dots & \dots \\ \dots & \sum_{a_j \neq a_{i'-1}} \theta_{a_{i'-1} a_j} & -\theta_{a_{i'-1} a_{i'+1}} & \dots \\ \dots & -\theta_{a_{i'+1} a_{i'-1}} & \sum_{a_j \neq a_{i'+1}} \theta_{a_{i'+1} a_j} & \dots \\ \dots & \dots & \dots & \dots \end{vmatrix} = D^{(i',i')}.$$

Объединяя полученные выражения для кратностей униграмм и для определителя, подставим их в формулу Теоремы 4 и получим:

$$N_\Theta = \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1 + \delta_{i'i})!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')}.$$

2) Пусть для любых i, j $\sum_{a_i \in A} \theta_{a_i a_j} = \sum_{a_i \in A} \theta_{a_j a_i}$. Это означает, что граф G_Θ является эйлеровым и первая буква совпадает с последней. Поскольку в эйлеровом цикле любая буква может быть первой (и соответственно последней), то нужно рассмотреть все варианты для каждой из букв a_i быть на последнем месте.

Пусть буква $a_{i'}$ — последняя (и соответственно первая) при прохождении эйлерова цикла. Тогда можно вычислить однозначно кратность любой униграммы θ_{a_i}. Если $a_i \neq a_{i'}$, то $\theta_{a_i} = \sum_{a_j \in A} \theta_{a_i a_j}$. При этом $\theta_{a_{i'}} = \sum_{a_j \in A} \theta_{a_{i'} a_j} + 1$. Таким образом, пользуясь результатами предыдущего пункта, получим число слов с данной матрицей кратностей биграмм Θ и последней буквой $a_{i'}$

$$N_{\Theta, a_{i'}} = \left(\sum_{a_j \in A} \theta_{a_{i'} a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')}.$$

Обратим внимание на то, что в данном случае можно вынести множитель $\sum_{a_j \in A} \theta_{a_{i'} a_j}$ из факториала, поскольку значение $\sum_{a_j \in A} \theta_{a_i a_j} - 1$, согласно условию теоремы, всегда неотрицательное и факториал определен.

Полное же число слов с данной матрицей кратностей биграмм Θ — это сумма величин $N_{\Theta, a_{i'}}$ по всем возможным последним буквам: $N_\Theta = \sum_{a_{i'} \in A} N_{\Theta, a_{i'}}$.

Поскольку соответствующий граф эйлеров, то согласно Лемме 1 для всех i, j $D^{(i,i)} = D^{(j,j)} = D$.

Собирая все воедино, получим искомую формулу:

$$N_\Theta = \sum_{a_{i'} \in A} N_{\Theta, a_{i'}} = \sum_{a_{i'} \in A} \left(\sum_{a_j \in A} \theta_{a_{i'} a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')} =$$

$$= \left(\sum_{a_i,a_j \in A} \theta_{a_i a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1)!}{\prod_{a_i,a_j \in A} \theta_{a_i a_j}!} D.$$

Найдем асимптотическую оценку для мощности языка согласно формулам предыдущей теоремы в случае матрицы кратностей биграмм $k\Theta$ при $k \to \infty$.

Определение 12. Матрица $\Theta \in \Xi$ называется положительной, если все элементы этой матрицы — натуральные числа.

Теорема 6. *Пусть задана положительная матрица Θ такая, что соответствующий ориентированный граф G_Θ эйлеров. Тогда при $k \to \infty$ для числа таких слов β_k, что $\Theta(\beta_k) = k\Theta$,*

$$N_{k\Theta} \sim c_2 * \frac{c_1^k}{k^{n(n-1)/2}},$$

где $c_1 = c_1(\Theta) > 1, c_2 = c_2(\Theta)$ — некоторые константы, зависящие только от изначальной матрицы Θ, а $n = |A|$ — мощность алфавита.

Доказательство. Поскольку ориентированный граф G_Θ эйлеров, можно воспользоваться п. 2 из формулировки Теоремы 5:

$$N_{k\Theta} = \left(\sum_{a_i,a_j \in A} k\theta_{a_i a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j} - 1)!}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})!} D_k =$$

$$= \left(\frac{\sum_{a_i,a_j \in A} k\theta_{a_i a_j}}{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})} \right) \left(\frac{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})!}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})!} \right) (D_k),$$

где D_k — главный минор матрицы Кирхгофа, построенной по матрице кратностей биграмм $k\Theta$.

Оценим по отдельности каждый из трех сомножителей в вышеприведенной формуле. Первый сомножитель имеет вид

$$\frac{\sum_{a_i,a_j \in A} k\theta_{a_i a_j}}{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})} = \frac{c_1' k}{c_2' k^n} = c_3' k^{1-n},$$

где c_1', c_2', c_3' — некоторые константы, зависящие только от матрицы Θ.

При переходе от Θ к $k\Theta$ все кратности биграммм умножаются на k, т.е. каждая из $n-1$ строк также умножается на k. Значит, по свойству определителя $D_k = k^{n-1} D$, при этом определитель D зависит только от Θ. Отсюда имеем, что

$$D_k = c_4' k^{n-1},$$

где c_4' — некоторая константа, зависящая только от матрицы Θ.

Для оценки последнего сомножителя воспользуемся формулой Стирлинга для факториала $\left(n! \sim \sqrt{2\pi n} (\frac{n}{e})^n\right)$:

$$\frac{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})!}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})!} \sim$$

$$\sim \frac{\prod_{a_i \in A} \sqrt{2\pi \sum_{a_j \in A} k\theta_{a_i a_j}} (\sum_{a_j \in A} k\theta_{a_i a_j})^{\sum_{a_j \in A} k\theta_{a_i a_j}} / e^{\sum_{a_j \in A} k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A} \sqrt{2\pi k\theta_{a_i a_j}} (k\theta_{a_i a_j})^{k\theta_{a_i a_j}} / e^{k\theta_{a_i a_j}}} =$$

$$= c_5' \frac{\prod_{a_i,a_j \in A} e^{k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A} e^{k\theta_{a_i a_j}}} \frac{k^{n/2}}{k^{n^2/2}} \frac{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})^{\sum_{a_j \in A} k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})^{k\theta_{a_i a_j}}} =$$

$$= c_5' \frac{1}{k^{(n^2-n)/2}} \frac{\prod_{a_i,a_j \in A} (\sum_{a_l \in A} k\theta_{a_i a_l})^{k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})^{k\theta_{a_i a_j}}},$$

где c_5' — некоторая константа, зависящая только от матрицы Θ.

Рассмотрим частное $\frac{1}{k\theta_{a_i a_j}} \sum_{a_l \in A} k\theta_{a_i a_l}$. Так как мы имеем положительную матрицу кратностей биграмм, то все кратности биграмм в любой ее строке отличны от нуля, поэтому, очевидно, имеет место $\frac{1}{k\theta_{a_i a_j}} \sum_{a_l \in A} k\theta_{a_i a_l} = 1 + \sigma_{ij}$, где σ_{ij} — некоторая положительная константа. Значит,

$$\frac{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})!}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})!} \sim c_5' \frac{1}{k^{(n^2-n)/2}} \prod_{a_i,a_j \in A} (1+\sigma_{ij})^{k\theta_{a_i a_j}} =$$

$$= c_5' \frac{1}{k^{(n^2-n)/2}} \left(\prod_{a_i,a_j \in A} (1+\sigma_{ij})^{\theta_{a_i a_j}} \right)^k = c_5' \frac{1}{k^{(n^2-n)/2}} c_1^k,$$

где c_1 — некоторая константа, зависящая только от матрицы Θ, при этом $c_1 > 1$ (как произведение положительных степеней чисел, больших единицы).

Соберем воедино:

$$N_{k\Theta} \sim c_3' k^{1-n} c_4' k^{n-1} c_5' \frac{1}{k^{(n^2-n)/2}} c_1^k = c_2 * \frac{c_1^k}{k^{n(n-1)/2}},$$

где c_2 — некоторая константа, зависящая только от матрицы Θ.

Замечание 3. Похожую оценку можно найти в работе [7], однако там авторы ограничились только верхней асимптотической оценкой.

Обозначим через Ξ_k множество матриц размера $|A| \times |A|$, каждый элемент которых представляет собой неотрицательное целое число, не превосходящее натуральное k. Все соотношения теперь будем рассматривать с учетом того, что все матрицы кратностей биграмм Θ принадлежат Ξ_k. Также будем считать, что исходный алфавит A зафиксирован и $|A| = n > 1$.

Замечание 4. Если две разные матрицы кратностей биграмм Θ_1 и Θ_2, $\Theta_1 \neq \Theta_2$, задают непустые языки $L(\Theta_1)$ и $L(\Theta_2)$, то очевидно, что $L(\Theta_1) \cap L(\Theta_2) = \varnothing$.

Через $FINITE(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих конечные (непустые) языки F_Θ.

Через $INFINITE(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные языки F_Θ.

Через $REG(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные регулярные языки F_Θ.

Через $NONREG(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные нерегулярные языки F_Θ.

Через $ALL(k)$ обозначим общее количество матриц $\Theta \in \Xi_k$.

Замечание 5. Очевидно, что $ALL(k) = (k+1)^{n^2}$.

Теорема 7. *С учетом введенных выше обозначений верны следующие соотношения*:

1) *для любого* k $\frac{1}{n(n-1)} < \frac{INFINITE(k)}{FINITE(k)} < 1$;
2) $\lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} = 0$;
3) $\lim_{k\to\infty} \frac{REG(k)}{NONREG(k)} = 0$.

Доказательство. 1) Для начала рассмотрим эйлеровы графы, отличные от одиночных (возможно, кратных) петель.

Заметим, что любой эйлеров граф превращается в полуэйлеров путем изъятия одного ребра, соединяющего различные вершины, из эйлерова графа. Это будет так по определению и свойству эйлеровых графов, так как при удалении одного ребра из эйлерова графа новых компонент связности не может возникнуть (в любом эйлеровом графе для создания двух и более компонент связности нужно удалить как минимум два ребра). Также, во всех вершинах, кроме тех двух, которые были началом и концом удаленного ребра, сумма входящих и сумма исходящих ребер не поменялась. При этом, если была удалено ребро $a_i \to a_j$, то в получившемся полуэйлеровом графе началом эйлерова пути будет a_j, а концом — a_i.

Необходимо заметить, что для двух разных эйлеровых графов полуэйлеровы графы, полученные такой операцией „удаления“ одного ребра, также будут разными. Также будут разными полуэйлеровы графы, полученные из одного эйлерова графа, если удалять ребра, соединяющие разные пары вершин. При этом каждому полуэйлерову графу будет соответствовать некоторый единственный эйлеров граф, получающийся добавлением одного ребра.

Обозначим через n_i^k число эйлеровых графов, построенных по матрицам из Ξ_k, в которых ровно i разных пар вершин соединены ребрами (пары упорядочены, т.е. (a_i, a_j) и (a_j, a_i) считаются разными парами). Тогда минимальное количество пар вершин для эйлерова графа равно 2 (минимальный цикл на двух вершинах), а максимальное — $n(n-1)$ (каждая пара вершин между собой соединена в двух противоположных направлениях). Каждому эйлерову графу с i парами соединенных вершин соответствует ровно i различных полуэйлеровых графов, которые получаются операцией „удаления“ одного ребра, соединяющего различные вершины эйлерова графа.

Значит, число эйлеровых графов без одиночных петель равно $\sum_{i=2}^{n(n-1)} n_i^k$, а общее число полуэйлеровых — $\sum_{i=2}^{n(n-1)} i n_i^k$.

Теперь рассмотрим эйлеровы графы, являющиеся одиночными петлями. Очевидно, из них операцией „удаления“ любого ребра (в данном случае — петли) мы не получим полуэйлерова графа. Всего различных одиночных (возможно, кратных) петель, построенных по матрицам из Ξ_k, будет равно количеству букв n в алфавите A, помноженному на максимальное количество k кратных петель в одной вершине, т.е. kn.

Таким образом, легко получить оценку снизу:

$$\frac{INFINITE(k)}{FINITE(k)} = \frac{\sum_{i=2}^{n(n-1)} n_i^k + nk}{\sum_{i=2}^{n(n-1)} i n_i^k} > \frac{\sum_{i=2}^{n(n-1)} n_i^k}{\sum_{i=2}^{n(n-1)} i n_i^k} \geqslant \frac{1}{n(n-1)}.$$

Для оценки сверху расчитаем величину n_2^k. Всего вариантов выбрать неупорядоченную пару различных вершин $\frac{n(n-1)}{2}$ (эйлеров цикл на двух вершинах содержит

как некоторое ребро $a_i \to a_j$, так и обязательно $a_j \to a_i$ при $i \neq j$). В каждой вершине независимо может быть от 0 до k петель, при этом эйлеров цикл на двух вершинах также может повторяться от 1 до k раз. Таким образом, $n_2^k = \frac{n(n-1)}{2} k(k+1)^2$.

Легко проверить, что для любых натуральных k будет верно соотношение $n_2^k + nk < 2n_2^k$. Поэтому верхняя оценка выводится следующим образом:

$$\frac{INFINITE(k)}{FINITE(k)} = \frac{\sum_{i=3}^{n(n-1)} n_i^k + n_2^k + nk}{\sum_{i=3}^{n(n-1)} i n_i^k + 2n_2^k} < \frac{\sum_{i=3}^{n(n-1)} n_i^k + 2n_2^k}{\sum_{i=3}^{n(n-1)} i n_i^k + 2n_2^k} \leqslant$$

$$\leqslant \frac{\sum_{i=3}^{n(n-1)} n_i^k + 2n_2^k}{3\sum_{i=3}^{n(n-1)} n_i^k + 2n_2^k} \leqslant 1.$$

В заключение заметим, что количество эйлеровых графов в точности соответствует количеству матриц кратностей биграмм, задающих счетные языки F_Θ, а количество полуэйлеровых графов — количеству матриц кратностей биграмм, задающих конечные (непустые) языки F_Θ.

2) Оценим сверху величину $INFINITE(k)$.

В эйлеровом графе количество входящих ребер всегда равно количеству выходящих для любой из n вершин — значит, можно записать систему линейных однородных уравнений для любого $i = 1, \ldots, n$:

$$\sum_{j=1}^{n} \theta_{a_i a_j} = \sum_{j=1}^{n} \theta_{a_j a_i}.$$

В этой системе n уравнений (по одному для каждой вершины), при этом n^2 неизвестных $\theta_{a_i a_j}$, $i, j = 1, \ldots, n$. Значит, размерность пространства решений данной системы линейных однородных уравнений равна разности размерности всего пространства переменных (n^2) и ранга матрицы, соответствующей этой системе (n).

Поскольку для любых $i, j = 1, \ldots, n$ верно $0 \leqslant \theta_{a_i a_j} \leqslant k$, то число решений есть не более чем произведение количества возможностей проварьировать каждую переменную ($k+1$ возможностей) для каждой из базисных переменных (которых по доказанному выше $n^2 - n$).

Количество решений системы не меньше количества эйлеровых графов (поскольку вышеприведенная система не учитывает связность графа), что, в свою очередь, равно $INFINITE(k)$. Также учтем, что $ALL(k) = (k+1)^{n^2}$. В итоге получим

$$\lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} \leqslant \lim_{k\to\infty} \frac{(k+1)^{n^2-n}}{(k+1)^{n^2}} = \lim_{k\to\infty} \frac{1}{(k+1)^n} = 0.$$

3) Согласно доказательству п. 2) Теоремы 4 из [4], множество матриц кратностей биграмм для регулярных языков F_Θ задает либо кратные одиночные петли, либо повторяющиеся циклы вида $a_{i_1} \to a_{i_2} \to \ldots \to a_{i_s} \to a_{i_1}$, где $i_{j_1} \neq i_{j_2}$ для $j_1 \neq j_2$, $s \leqslant n$.

Всего одиночных некратных петель равно числу букв алфавита A, т.е. n штук. Число простых неповторяющихся циклов длины $s \geqslant 2$ есть также функция от n и не зависит от k. Таким образом, суммарное количество одиночных некратных петель и простых неповторяющихся циклов любой длины от 2 до n есть некая функция $f(n)$, не зависящая от k.

Каждые из таких одиночных некратных петель или простых неповторяющихся циклов мы можем изменять следующим образом, оставаясь в рамках Ξ_k: умножать

количество петель (или количество ребер для любых двух соединенных ребром вершин) на одно и то же число k_1, где $1 \leqslant k_1 \leqslant k$. В итоге общее число матриц из Ξ_k, задающих регулярные языки F_Θ, есть $REG(k) = kf(n)$.

Для оценки числа матриц из Ξ_k, задающих нерегулярные языки F_Θ, возьмем любый простой цикл, содержащий все n вершин, например, $a_1 \to a_2 \to ... \to a_n \to a_1$. Для того, чтобы перевести данный эйлеров граф в разряд графов, задающих нерегулярные языки, достаточно добавить в какую-либо из вершин петлю, возможно, кратную. Всего возможностей разместить петли в вершинах такого графа будет $(k+1)^n - 1 > k^n$ для $n > 1$.

Всего же матриц из Ξ_k, задающих нерегулярные языки F_Θ, будет очевидно не меньше, чем таких простых циклов с петлями. Значит, $NONREG(k) > k^n$, и

$$\lim_{k\to\infty} \frac{REG(k)}{NONREG(k)} \leqslant \lim_{k\to\infty} \frac{kf(n)}{k^n} = \lim_{k\to\infty} \frac{f(n)}{k^{n-1}} = 0.$$

Следствие 1. *Если обозначить через $NONEMPTY(k)$ количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих непустые (как конечные, так и счетные) языки F_Θ, то*

$$\lim_{k\to\infty} \frac{NONEMPTY(k)}{ALL(k)} = 0.$$

Доказательство. Согласно п. 1) Теоремы 7 $INFINITE(k)$ и $FINITE(k)$ имеют одинаковые порядки роста по k. Так как согласно п. 2) $\lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} = 0$, то и $\lim_{k\to\infty} \frac{FINITE(k)}{ALL(k)} = 0$. С учетом того, что $NONEMPTY(k) = INFINITE(k) + FINITE(k)$, получим

$$\lim_{k\to\infty} \frac{NONEMPTY(k)}{ALL(k)} = \lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} + \lim_{k\to\infty} \frac{FINITE(k)}{ALL(k)} = 0.$$

Работа выполнена на кафедре МаТИС МГУ им. М. В. Ломоносова под руководством проф. Д. Н. Бабина.

Список литературы

1. Марков А.А., “Пример статистического исследования над текстом „Евгения Онегина“, иллюстрирующий связь испытаний в цепь”, *Известия Императорской Академии наук*, 6, **7**, 1913, 153–162.
2. Essen U., Steinbiss V., “Cooccurrence smoothing for stochastic language modeling”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **1** (1992), 161–164.
3. Hutchinson J. P., Wilf H. S., “On eulerian circuits and words with prescribed adjacency patterns”, *J. Comb. Theory*, **18** (1975), 80–87.
4. Петюшко А. А., “О биграммных языках”, *Дискретная математика*, **25**:3 (2013), 64–77.
5. Оре О., *Теория графов*, Наука, Москва, 1980.
6. Chung F. R. K., *Spectral graph theory*, RI: Amer. Math. Soc., Providence, 1997.
7. Kim K.H., Roush F., “Words with prescribed adjacencies”, *J. Comb. Theory*, **26**:1 (1979), 85–97.

Статья поступила 09.11.2012.