Классификация Хомского для матриц биграммных языков

А. А. Петюшко, Д. Н. Бабин

Множество слов, у которых частоты встречаемости пар соседних букв образуют одну и ту же матрицу - это формальный (биграммный) язык. В статье описывается матрицы, соответствующие регулярным и контекстно-свободным биграммным языкам.

Ключевые слова: биграммный язык, матрица частот, соседние буквы, эйлеровы циклы.

Матрицы традиционно являются инструментом решения прикладных задач. Ещё в начале 20 века выдающимся русским учёным А. А. Марковым [1] был создан аппарат цепей, впоследствии названных цепями Маркова. В дальнейшем этот аппарат получил широкое применение для распознавания и статистического моделирования естественных языков, погоды и других явлений [2]. При таком подходе целью является сама матрица, по которой можно будет предсказывать дальнейшее поведение изучаемого объекта.

В данной статье изучается обратная задача: рассматривается формальный язык с фиксированной матрицей биграмм. В классификации Хомского [3] принадлежность слова языку определяется либо с помощью автомата (регулярные языки), либо с помощью автомата с магазинной памятью (контекстно-свободные языки), либо с помощью машины Тьюринга [4]. Нашей целью является установление соответствия между матричным описанием языка и описанием Хомского.

Пусть A ($|A| < \infty$) - конечный алфавит. *Биграммой* в алфавите A называется двухбуквенное слово $ab \in A^*, a, b \in A$

($ab \neq ba$ при $a \neq b$). Обозначим через $\theta_\beta(\alpha)$, где $\beta \in A^*, \alpha \in A^*$, количество подслов β в слове α, т.е. количество различных разложений слова α в виде $\alpha = \alpha'\beta\alpha''$ (α' и α'' могут быть пустыми).

По слову $\alpha \in A^*$ можно построить квадратную матрицу биграмм $\Theta(\alpha)$ размера $|A| \times |A|$ такую, что на месте (i, j) матрицы будет стоять значение $\theta_{a_i a_j}(\alpha)$. Обозначим через Ξ множество квадратных матриц размера $|A| \times |A|$, каждый элемент которых является неотрицательным целым числом.

Назовём *языком* $L(\Theta)$*, порожденным матрицей* $\Theta \in \Xi$, множество всех слов, имеющих одну и ту же матрицу биграмм Θ,

$$L(\Theta) = \{\beta \in A^* | \Theta(\beta) = \Theta\}.$$

Построим по матрице Θ ориентированный граф G_Θ, вершинами которого будут все буквы из алфавита A, при этом ребра будут соответствовать биграммам с учётом их кратностей, т.е. кратность θ_{ab} будет порождать θ_{ab} ориентированных рёбер $a \to b$. Аналогично, кратность θ_{cc} будет порождать θ_{cc} петель $c \to c$.

Ориентированный граф назовём *эйлеровым*, если выполняются следующие условия:

1) Неизолированные вершины составляют одну компоненту связности соответствующего неориентированного графа;

2) У всех вершин количество входящих рёбер равно количеству исходящих рёбер.

Назовём ориентированный граф *почти эйлеровым*, если выполняется условие 1), и у всех вершин, кроме двух, количество входящих рёбер равно количеству исходящих рёбер, а у оставшихся двух вершин разность количества входящих рёбер и количества исходящих рёбер равна +1 и -1, соответственно.

Известно [5], что в эйлеровом графе существует эйлеров цикл, а в почти эйлеровом - эйлеров путь, не являющийся эйлеровым циклом (то есть начальная вершина пути не совпадает с конечной).

Имеет место следующий критерий непустоты:

Теорема 1. [6] *Для того, чтобы язык $L(\Theta)$ был непуст, достаточно, чтобы построенный по Θ ориентированный граф G_Θ был либо эйлеровым, либо почти эйлеровым.*

Следствие 1. *Проблема пустоты языка $L(\Theta)$ алгоритмически разрешима.*

Более интересен случай, когда матрица биграмм рассматривается как матрица пропорций биграмм, то есть языка, в котором сохраняются отношения $\theta_{ab}(\alpha)/\theta_{cd}(\alpha)\ \forall a,b,c,d \in A$, $\theta_{cd}(\alpha) > 0$ для любого слова α из этого языка.

Назовём *биграммным языком, заданным матрицей $\Theta \in \Xi$,* назовём

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

Имеет место теорема о числе слов в частотных языках:

Теорема 2. [6] *Пусть задан набор биграмм $\Theta \in \Xi$. Тогда*
1) Если ориентированный граф G_Θ является эйлеровым, то в частотном языке F_Θ счетное число слов;
2) Если ориентированный граф G_Θ является почти эйлеровым, то частотный язык F_Θ совпадает с $L(\Theta)$ (и, соответственно, в нем конечное число слов);
3) Если ориентированный граф G_Θ не является ни эйлеровым, ни почти эйлеровым, то в частотном языке F_Θ нет ни одного слова.

Имеет место

Следствие 2. *Задача определения по набору значений $\Theta \in \Xi$, пуст, конечен или счетен частотный язык F_Θ, алгоритмически разрешима.*

Назовём две ненулевые матрицы Θ_1 и Θ_2 из Ξ *кратными*, если существует действительный коэффициент $c \in R, c \neq 0$, такой что верно $\Theta_1 = c\Theta_2$. В противном случае ненулевые матрицы назовём некратными.

Теорема 3. [6] *Пусть $A, |A| < \infty$ - некоторый конечный алфавит, а матрица биграмм Θ такова, что ориентированный граф G_Θ является эйлеровым. Тогда:*
1) Если существует такое разложение $\Theta = \Theta_1 + \Theta_2$ в сумму

двух ненулевых некратных матриц, причём обе матрицы Θ_1 *и* Θ_2 *задают ориентированные графы* G_{Θ_1} *и* G_{Θ_2}*, которые являются эйлеровыми, то язык* F_Θ *нерегулярен;*
2) В противном случае язык F_Θ *регулярен.*

Обозначим через $\Xi_k \subset \Xi$ множество матриц размера $|A| \times |A|$, каждый элемент которых не превосходит $k > 0$. Через N_f^k - количество тех из них, которые задают конечные (непустые) языки F_Θ, через N_i^k - количество матриц, задающих счётные языки F_Θ, через N_{reg}^k - количество матриц, задающих счётные регулярные языки F_Θ, через N_{nreg}^k - количество матриц, задающих счётные нерегулярные языки F_Θ. Через N^k обозначим общее количество матриц биграмм $\Theta \in \Xi_k$. Очевидно, что $N^k = (k+1)^{n^2}$.

Теорема 4. [7] *Выполнены следующие соотношения:*
1) Для всех $k > k_0$ *выполнено* $\frac{1}{n(n-1)} < \frac{N_i^k}{N_f^k} < 1$*;*
2) $\lim_{k\to\infty} \frac{N_i^k}{N^k} = 0$*;*
3) $\lim_{k\to\infty} \frac{N_{reg}^k}{N_{nreg}^k} = 0$.

Обозначим через N_q^k количество матриц биграмм $\Theta \in \Xi_k$, задающих непустые языки F_Θ. Имеет место следствие:

Следствие 3. $\lim_{k\to\infty} \frac{N_q^k}{N^k} = 0$.

Имеет место критерий контекстно-свободности биграммного языка:

Теорема 5. [8] *Пусть матрица кратностей биграмм* $\Theta \in \Xi$*, задающая эйлеров граф, разлагается в сумму не менее двух линейно независимых матриц, также задающих эйлеровы графы. Тогда:*
1) *Если* Θ *разлагается* ***единственным образом*** *в сумму двух линейно независимых матриц* $\Theta = \Theta_1 + \Theta_2$*, соответствующих* ***простым*** *эйлеровым циклам, и не разлагается в сумму большего количества линейно независимых матриц, соответствующих эйлеровым циклам, то язык* F_Θ *— контекстно-свободный;*
2) *Иначе язык* F_Θ *— не контекстно-свободный.*

Список литературы

[1] А. А. Марков. Пример статистического исследования над текстом “Евгения Онегина”, иллюстрирующий связь испытаний в цепь. // Известия Императорской Академии наук, серия VI. – 1913. – Т. 10. – № 3. – С. 153–162.

[2] U. Essen and V. Steinbiss. Cooccurrence smoothing for stochastic language modeling. // Acoustics, Speech, and Signal Processing, International Conference on. – IEEE, 1992. – Vol. 1. – P. 161–164.

[3] N. Chomsky. Three models for the description of language. // Information Theory, IRE Transactions on. – 1956. – Vol. 2. – № 3. – P. 113–124.

[4] В. Б. Кудрявцев, С. В. Алешин, А. С. Подколзин. Введение в теорию автоматов. – М.: Наука, 1985.

[5] О. Оре. – М.: Наука, 1980.

[6] А. А. Петюшко. О биграммных языках. // Дискретная математика. – 2013. – Т. 25. – № 3. – С. 64–77.

[7] А. А. Петюшко. О мощности биграммных языков. // Дискретная математика. – 2014. – Т. 26. – № 2. – С. 71–82.

[8] А. А. Петюшко. О контекстно-свободных биграммных языках. // Интеллектуальные системы. Теория и приложения. – 2015. – Т. 19. – № 2. – С. 187–208.