МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1967 год Выпуск 88

УДК 655.25+801.4

АВТОМАТИЧЕСКОЕ ВЫЯВЛЕНИЕ СТРУКТУРЫ ПАРАДИГМАТИЧЕСКИХ ОТНОШЕНИЙ ТЕКСТА

М.К.Тимофеева

I. Известно, что каждая единица языковой системы противопоставляется некоторой другой единице, вступая с ней в парадигматические отношения. В связи с этим представляется интересным автоматическое построение формальной модели парадигматики произвольного текста, представляющего некоторый язык. Предлагаемый нами метод использует наиболее универсальные свойства известных языковых систем и состоит в автоматическом выявлении конкретных средств реализации этих свойств в предложенных текстах. Возможность полной автоматизации процесса обусловлена выбором таких универсальных свойств, которые опираются не на семантику текста, а на его статистико-комбинаторные характеристики. Это, в свою очередь, приводит к тому, что выявляются не все, а только наиболее типичные свойства парадигматики текста. Степень адекватности модели действительной системе парадигматических отношений языка зависит от величины и информативности текста, его представляющего.

Итеративное применение метода позволяет строить иерархию отношений единиц текста. Исходный материал для первой итерации - 1-граммный спектр текста [I] (частотный словарь встречаемости в нем подцепочек различной длины), для каждой последующей - 1-граммный спектр текста, полученного из предыдущего путем замены каждой выделенной в нем единицы кодом соответствующей ей парадигмы.

Предлагаемый метод анализа текста может быть полезен при разработке формы представления текста для работы со стенорайтерами, при автоматизации редакционно-издательских работ, при сжатии информации, хранимой в памяти ЭВМ, а также при типологическом изучении языков и выявлении тенденций их развития.

2. Известен ряд работ [2,3], предлагающих формальные методы выделения значимых подцепочек текста на основе использования универсальных свойств языков.

В [2] выделение значимых подцепочек текста производится с помощью анализа спектра 1-грамм, но задача установления парадигматических отношений между этими подцепочками не ставится. Выделенное множество значительно шире множества грамматических единиц текста. При больших объемах текстов производить отбор таких единиц вручную было бы затруднительно. Для автоматизации этого процесса необходимо ввести дополнительные критерии выделения грамматических единиц, учитывающие отношения между ними.

В [3] ставится задача выявления парадигматических отношений, но накладываются ограничения на допустимые позиции единиц и их длину. Основным критерием выделения единиц служит предположение о том, что в каждой парадигме найдется хотя бы одна единица, для которой условная вероятность встречаемости в тексте (в заданной позиции или в заданном окружении) существенно отличается от независимой. Введение такого критерия значительно уменьшает перебор подцепочек текста в ходе выделения единиц, но при снятии ограничений на позицию и длину этих единиц его использование становится трудоемким. Кроме того, добавление еще одного критерия, не связанного с выявлением отношений, несколько сужает область применения метода, так как указанное предположение может выполняться не для всех языков.

Благодаря использованию закономерности, обнаруженной на спектрах 1-грамм, появилась возможность достаточно быстрой организации перебора подцепочек текста. Основной критерий выделения единиц - наличие парадигматических отношений между ними. Остальные критерии служат для отбрасывания тех единиц, которые в традиционной лингвистике обычно не рассматриваются как самостоятельные. После окончания работы алгоритма нахождение отброшенных единиц, если это необходимо, не представляет затруднений.

3. Рассмотрим свойства языковых систем, положенные в основу метода. Пусть задан текст T в алфавите A, через A^* обозначим множество всех цепочек, состоящих из символов алфавита A; Λ - пустая цепочка, W - множество всех подцепочек T, обладающих заданным набором формальных признаков. Причем W должно быть таким, чтобы для любых двух $\omega_1, \omega_2 \in W$ не существовало непустых $g, \omega_1', \omega_2' \in A^*$ таких, что $\omega_1 = \omega_1' g$, $\omega_2 = g\omega_2'$ и $\omega_1' g \omega_2'$ - подцепочка T.

Множество $q_{f_i,f_j} \in W$ образует к в а д р а т[*] для $f_i, f_j \in A^*$, если

1) $f_i \neq f_j$, $f_i \cdot f_j \neq \Lambda$;

2) найдутся такие непустые и несовпадающие цепочки, $\Delta_1\gamma_1, \Delta_2\gamma_2 \in A^*$, что q_{f_i,f_j} представимо в виде:

$$q_{f_i,f_j} = \{ \Delta_1 f_i \gamma_1, \Delta_1 f_j \gamma_1, \Delta_2 f_i \gamma_2, \Delta_2 f_j \gamma_2 \}.$$

Упорядоченные пары $(\Delta_1,\gamma_1), (\Delta_2,\gamma_2)$ - к о н т е к с т ы для f_i, f_j; $K(f_i,f_j)$ - множество всех контекстов для f_i, f_j; $K(f_i) = \cup_j K(f_i,f_j)$ - множество контекстов для f_i; $k_{f_i,f_j} = |K(f_i,f_j)|$; S - некоторая заданная константа больше 2.

Цепочки $f_i, f_j \in A^*$ в з а и м о з а м е н я е м ы, если $k_{f_i,f_j} \geq S$.

Подцепочка f_i текста T у ч а с т в у е т в с и с т е м е п а р а д и г м а т и ч е с к и х о т н о ш е н и й, если найдется цепочка f_j, с которой она взаимозаменяема. Обозначим через F множество всех таких подцепочек текста T и через $P_1, \ldots, P_n$ - максимальные множества попарно взаимозаменяемых элементов из F.

Для каждого $f_i \in F$ упорядочим некоторым образом контексты $K(f_i)$. В х о ж д е н и е f_i - это пара (f_i, N), где N - номер контекста из $K(f_i)$.

Рассмотрим такие вхождения $(f_i,N_1), (f_j,N_2)$ разных цепочек f_i, f_j, для которых $\Delta_1 f_i \gamma_1 = \Delta_2 f_j \gamma_2$, где (Δ_1, γ_1) - N_1-й контекст из $K(f_i)$, (Δ_2,γ_2) - N_2 -й контекст из $K(f_j)$, $n_1 = |\Delta_1|$, $n_2 = |\Delta_1 f_i|$, $n_3 = |\Delta_2|$, $n_4 = |\Delta_2 f_j|$. Вхождение (f_i,N_1) в л о ж е н о во вхождение (f_j,N_2), если $n_1 \geq n_3$, $n_2 \leq n_4$. Вхождения (f_i,N_1) и (f_j,N_2) п е р е с е к а ю т с я, если либо $n_3 \leq n_1 < n_4$, либо $n_3 < n_2 \leq n_4$.

Элемент $f_i \in F$ м и н и м а л е н[**], если для любого P_r, содержащего f_i, не существует таких цепочек $g,h \in A^*$ $(g \cdot h \neq \Lambda)$, что любое $f_j \in P_r$ представимо в виде $f_j = g f'_j h$, $f'_j \in A^*$.

Далее, обозначим через $K(P_r)$ множество контекстов, поставленное в соответствие множеству P_r и состоящее из таких контекстов (Δ_1, γ_1), для которых существует хотя бы одно $f_i \in P_r$, такое, что $(\Delta_1, \gamma_1) \in K(f_i)$ и f_i не входит ни в одно из P_j, $j \neq r$.

---

[*] Это понятие опирается на обобщенный "квадрат Гринберга" [4].

[**] Сходные свойства минимальности используются в [3].

Контексты множества $K(f_i)$ минимальны, если для любого P_r, содержащего f_i, не существует таких цепочек $g, h \in A^*(g \cdot h \neq \Lambda)$, что каждый контекст $(\Delta_1, \Upsilon_1) \in K(P_r)$ представим в виде $(\Delta_1' g, h \Upsilon_1')$, где $\Delta_1', \Upsilon_1' \in A$ и элементы множества $P_r' = \{g f_i h \mid f_i \in P_r\}$ минимальны.

Грамматические единицы - минимальные вступающие в парадигматические отношения подцепочки текста T, вхождения которых не пересекаются и контексты минимальны.

Парадигмы - максимальные множества попарно взаимозаменяемых грамматических единиц текста T.

Омонимы - совпадающие по написанию грамматические единицы, входящие в разные парадигмы.

Качество выделения грамматических единиц зависит от выбора W и S. В первом варианте алгоритма предполагается, что в алфавите A выделено некоторое подмножество разделителей R, содержащее пробел и знаки препинания. Параметр S рассматривается как равный 2. Множество W состоит из всех подцепочек T, расположенных в нем между двумя символами из R и не содержащих их внутри себя. Если R неизвестно, то при выборе W используются спектры 1-грамм текста T. Для выработки критериев построения W в общем случае эксперименты проводятся на спектрах 1-грамм, хотя при известном R можно было бы использовать более простые данные. Алгоритмы анализа элементов W не зависят от способа его построения.

4. Анализ текста осуществляется посредством итеративной процедуры, на каждом шаге которой выделяются только те грамматические единицы, для которых не существует вложенных вхождений других грамматических единиц. Кодируются некоторым образом парадигмы, обнаруженные на каждом шаге. Производится обращение к алгоритму выявления минимальных общих признаков элементов каждого множества $K(P_r)$, отличающих их от элементов любого другого множества $K(P_j)$ [6]. Множества $K(P_r)$ дополняются контекстами из $K(f_i)$, которые по выделенным признакам могут быть однозначно отнесены к какой-либо парадигме. Все $\Delta_1 f_i \Upsilon_1, (\Delta_1, \Upsilon_1) \in K(P_r)$ в тексте заменяются на цепочки $\Delta_1 p_r \Upsilon_1$, где p_r - код парадигмы P_r. Указанное свойство элементов множества W обеспечивает однозначность такой перекодировки. На следующей итерации анализируется перекодированный текст.

Спектр 1-грамм текста и элементы W упаковываются в виде деревьев, построение которых можно иллюстрировать следующим примером.

Задан спектр 1-грамм:

l = 4 ␣ ПРИ - 559

l = 5 ␣ ПРИМ - I74; ␣ ПРИ - I06

l = 6 ␣ ПРИМЕ - I7I

l = 7 ␣ ПРИМЕР - 96; ␣ ПРИМЕН - 69

l = 8 ␣ ПРИМЕРЕ - I3; ␣ ПРИМЕР. - 5; ␣ ПРИМЕРЫ - 8; ␣ ПРИМЕНЕ - 37

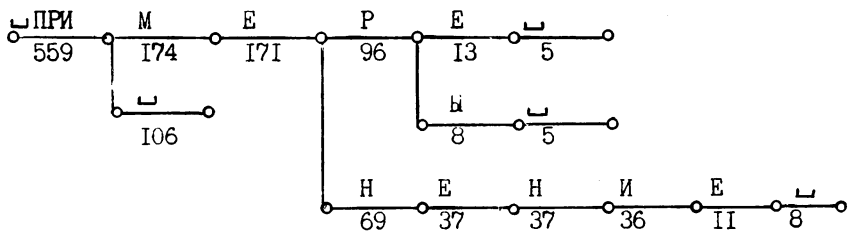
l = 9 ␣ ПРИМЕРЕ␣ - 5; ␣ ПРИМЕРЫ, - 4; ␣ ПРИМЕНЕН - 37

l = I0 ␣ ПРИМЕНЕНИ - 36

l = II ␣ ПРИМЕНЕНИЕ - II

l = I2 ␣ ПРИМЕНЕНИЕ ␣ - 8

Ему ставим в соответствие дерево:



Если в результате анализа такого дерева грамматические единицы не выявлены, то рассматривается обратное дерево, в котором элементы W упакованы не с начала, как в данном, а с конца.

Звено дерева - путь ρ, образованный вершинами $v_1, \ldots, v_m$, в котором $m \geq 2$ и

I) из v_1 выходит хотя бы одна дуга, ведущая не в v_2;

2) если из v_r $(1 < r < m)$ выходит только одна дуга, то путь, образованный вершинами $v_r, \ldots, v_m$, - линейный;

3) путь ρ не может быть продолжен без потери двух первых свойств.

Используются следующие способы ускорения просмотра деревьев в ходе выявления грамматических единиц.

I) Так как $S \geq 2$, то достаточно рассматривать только те контексты, частота которых не менее двух.

2) В силу требования минимальности грамматических единиц каждая парадигма P_r содержит хотя бы два элемента f_i и f_j, первые символы которых различны. Такие элементы выявляются в первую очередь с помощью анализа точек ветвления дерева. При этом будут обнаружены все остальные элементы парадигмы (так как они взаимозаменяемы с f_i, f_j), но будут выявлены не все контексты, в которых

они взаимозаменяемы друг с другом. Отношения между этими элемен - тами парадигмы P_r устанавливаются с помощью анализа их вхождений, не рассмотренных при выявлении f_i, f_j.

3) Шаг итерации состоит из нескольких этапов, на каждом из которых анализируются пути, содержащие r звеньев ($r \in \{1,2, \ldots, r_{max}\}$, где r_{max} - максимальное число звеньев, встречающихся на одном пути в дереве). Первоначально r=1. Пути, на которых выяв - лены грамматические единицы, на данном шаге больше не рассматри - ваются.

Причиной такого разбиения дерева послужила закономерность, согласно которой границы между значимыми единицами языка часто совпадают с границами между звеньями [5]. Единицы, для которых эта закономерность не выполняется, выявляются во вторую очередь. При этом число путей дерева, подлежащих анализу, существенно меньше первоначального.

4) Множество W разбивается на части в зависимости от начальных символов содержащихся в нем подцепочек. Каждая часть упаковывается в отдельное дерево. В результате анализа одного из деревьев, содержащих максимальное число вершин, выделяются некоторые множества грамматических единиц и парадигм. Рассматривается следующее по числу вершин дерево, в результате чего полученные множества дополняются. Анализ деревьев прекращается, как только добавление нового дерева не приводит к изменению множеств грамматических единиц и парадигм. Таким образом, если для каждого P_r пересечение множеств $K(f_i, f_j)$, где $f_i, f_j \in P_r$, велико, то доста - точно рассмотреть небольшое число деревьев 1-грамм.

5. Реализована первая часть алгоритма, состоящая из построения деревьев 1-грамм и выявления грамматических единиц. Исполь - зовались спектры 1-грамм для технического текста на русском языке (см. [1]). Проведен один шаг итерации, на котором рассматривались пути, содержащие одно звено, и такие контексты (Δ_1, γ_1), в которых $\gamma_1 = \Lambda$.

В результате анализа одного дерева выделены звенья:

ПОЛОЖИТЕЛЬН:ОГО,ОЕ,ОЙ,ЫМИ,ЫХ
ПОЛУЧЕН:НОГО,НЫЕ,НЫЙ
ПОТЕНЦИАЛ:А, Л ,ЬНОЙ,ОВ
ПОВЕРХНОСТ:И,Ь,НОЙ,НАЯ,НЫЙ,НЫХ,ЯМИ,ЯМ
ПОСТОЯН:НО,НЫМИ,НЫХ,НЫЕ,СТВА
ПРОИЗВОЛЬН:УЮ,ОГО,ОЙ,ОЕ,ЫХ,ЫЕ,ЫМ
ПОЛИНОМ:ОВ,АМ
ПОКАЗАТЕЛ:Е,Ь
ПОПЕРЕЧН:ОГО,ОЕ
ПЕРЕПОЛНЕНИ:Я,Е
ПЕРЕНОС:А,Е
ПЕРЕМЕННН:ОЙ,ОГО

ПРОСПЕКТИВН:ЫЕ,ЫХ,АЯ,ОЙ
ПРОСТРАНСТВ:А,О,Е,ЕННЫХ
ПРОПОРЦИОНАЛЬН:А,О,ОСТИ
ПРЕДЫДУЩ:ИХ,ЕГО,ЕМ
ПРЕДСТАВ:ЛЯЕТ,ЛЯТЬ,ЛЕНО,ИТЬ,ИМЫХ
ПРЕОБРАЗОВАНИ:Й,ЕМ
ПАРАЛЛЕЛЬН:ЫХ,ОГО
ПРОВОДНИК:ОМ,А, Л,
ПРАВИЛЬН:ОЙ,ЫХ
ПРАКТИЧЕСК:ОЙ,ОЕ
ПРИНЦИП:А,ОМ
ПРОИЗВЕДЕНИ:Я,Е,Й
ПРОЦЕСС:А,Е
ПЛОТНОСТ:И,ЬЮ,Ь,ЕЙ
ПЛОСКОСТ:И,Ь,ЯМИ
ПАРАГРАФ:Е,АХ
ПОСЛЕДОВАТЕЛЬ:НОЕ,НОСТЬ,НОСТИ,НО,НОГО,НЫМ,НЫХ.

Первая цепочка символов, отделенная двоеточием,соответствует пути дерева,ведущему из его корня в первую вершину звена, цепочки после двоеточия - полным путям поддерева с корнем в этой вершине. При анализе дерева выделены следующие грамматические единицы: АЯ, ОЙ, ОЕ, ОГО, ОМ, ЫМ, ЫЕ, ЫХ, ЫМИ, А, И, Ь, Я, О, Е, ЯМИ, Л, ОСТИ. Из них неверно выделена только последняя единица, эта ошибка объясняется тем, что проверка вложенности единиц в данном эксперименте не проводилась. Часть грамматических единиц осталась невыявленной, так как использовался спектр 1-грамм,содержащий только 1-граммы неединичной частоты.

Рассмотрим основные типы ошибок в выделении окончаний и их парадигм:

I) Словам, относящимся к двум разным синтаксическим классам, но имеющим одну основу,сопоставляется одна парадигма вместо двух. Например, основе ПОСЛЕДОВАТЕЛЬН- как краткому прилагательному соответствует парадигма {О,А}, а как полному прилагательному - {ОЕ, ОГО,ОМУ,...}. Алгоритмом будет выделена одна парадигма {О,А, ОЕ, ОГО,ОМУ,...}.

2) Если имеется чередование символов в основе (в том числе и чередование с пустой цепочкой), то граница, отделяющая основу от окончания, будет смещена. Например, в словах ТОЧЕН, ТОЧНА, ТОЧНО выделяется окончания ЕН, НА, НО вместо Л, А, О.

Большая часть ошибок указанных типов будет устранена в следующем варианте алгоритма при дополнении модели синтагматическими отношениями текста (т.е. отношениями между соседними его единицами), также выявляемыми автоматически, и использовании более тонких признаков парадигм.

Алгоритмы реализованы на RTK ОС ЕС.

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. СУХОТИН Б.В. Оптимизационные методы исследования языка.- М.: Наука, 1976. - 120 с.

3. АНДРЕЕВ Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. -Л.: Наука, 1967. - 403 с.

4. ГРИНБЕРГ Дж. Квантитативный подход к морфологической типологии языков. - В кн.: Новое в лингвистике. Вып. 3. М., 1963, с. 60-95.

5. ГУСЕВ В.Д., КОСАРЕВ Ю.Г., ТИТКОВА Т.Н. Методы поиска и анализ статистических закономерностей в символьных последовательностях. -В кн.: Машинные методы обнаружения закономерностей. (Материалы Всесоюз. симпозиума, 5-7 апреля 1976). Новосибирск, 1976, с. 75-84.

6. КОСАРЕВ Ю.Г., ЧУЖАНОВА Н.А. Автоматический синтез алгоритмов классификации словоформ по типам словоизменительных парадигм. -В кн.: Структурная и математическая лингвистика. Киев, 1978, с.24-32.

Поступила в ред.-изд.отд.
29 октября 1980 года