Знания-Онтологии-Теории (ЗОНТ-09)

Методы выделения структурных единиц в символьных последовательностях. Межъязыковые аналогии.

Гусев В.Д. [1], Мирошниченко Л.А.[1], Саломатина Н.В.[1]

[1]*Институт Математики им. С.Л. Соболева СО РАН, пр. ак. Коптюга , д.4, г.Новосибирск, 630090, Россия.*

gusev@math.nsc.ru, luba@math.nsc.ru, nataly@math.nsc.ru

Аннотация. Исследуются возможности формализации понятия "структурная единица" в различных языковых системах с учётом таких её атрибутов как повторяемость, вариативность, иерархичность. Сделан вывод о том, что универсальной базой для разработки алгоритмов выделения структурных единиц нижнего уровня могут служить *L*-граммные характеристики текста в сочетании с позиционной информацией о местах вхождения каждой *L*-граммы в текст. Многие подходы доведены до алгоритмов и апробированы на реальных данных. Показана роль межъязыковых аналогий в постановке и решении задач данного типа.
Ключевые слова: символьные последовательности, тексты, структурные единицы, сегментация, сложностной анализ, *L*-граммный анализ, межъязыковые аналогии

1 Введение

Построение онтологий различных предметных областей в существенной мере опирается на информацию, представленную в символьной форме. Сюда относятся:
- тексты на естественном языке, описывающие данную предметную область;
- символьные последовательности, являющиеся объектом изучения во многих предметных областях (ДНК- и аминокислотные последовательности, тексты программ, шифротексты, музыкальные произведения и др.).

Тексты на естественном языке используются для выделения терминов предметной области и взаимосвязей между ними. Эти тексты обычно являются структурированными в том смысле, что в них с помощью разделителей и других знаков пунктуации выделены слова, предложения абзацы и иные элементы структуры. Однако часто возникает необходимость в выделении элементов промежуточных уровней иерархии, например, устойчивых словосочетаний, типичных для терминологической лексики, или сверхфразовых единств, характеризующих макроструктуру текста. Для выделения таких структурных единиц (СЕ) требуется разрабатывать специальные алгоритмы.

Проблема выделения структурных единиц становится ещё более актуальной при работе с символьными последовательностями, не содержащими знаков пунктуации, или содержащими их в неявной (скрытой) форме (иероглифическое письмо, биологические макромолекулы, знаменные песнопения и т.п.). Некоторые объекты, например, геномы различных организмов, описываются последовательностями длиной в миллионы и даже миллиарды символов. Работа с ними требует предварительного расчленения их на более мелкие сегменты в соответствии с

определёнными критериями. Процедура разбиения носит, как правило, иерархический характер. Элементы разбиений трактуются как структурные единицы различных иерархических уровней.

Целью работы является исследование возможностей *формализации понятия* "*структурная единица*" с учётом таких её атрибутов как *повторяемость*, *вариативность*, *иерархичность* и разработка *алгоритмов выделения структурных единиц*, легко настраиваемых на конкретную предметную область. *Специфика нашего подхода* определяется использованием *межъязыковых аналогий*, *максимально широкой трактовкой понятия повтора*, являющегося основным структурообразующим элементом текста, *учётом* для оценки значимости выделяемых структурных единиц не только частотной, но и *позиционной информации*.

Основное внимание будет уделено *нижним уровням* языковой иерархии. Методы, используемые для крупноблочной сегментации, относятся уже к текстовому уровню и обычно сводятся к решению задачи о разладке категориальных данных, т.е. к поиску точек "значимого" изменения статистических свойств анализируемой последовательности ("multiple change-point problem"). Обзор соответствующих методов сегментации представлен в [1]. Отметим, в частности, метод максимального правдоподобия [2], байесовское приближение [3], скрытые марковские модели [4] и (самое простое) сравнение распределений элементов алфавита в полуокнах, расположенных слева и справа от потенциально возможной точки сегментации с использованием, например, дивергенции Йенсена-Шеннона [5].

Работ по выделению структурных единиц нижнего уровня не так много (укажем для примера [6÷8]). Это объясняется тем, что часть из них уже зафиксирована в словарях, составленных вручную. Упомянутые выше статистические методы на нижних уровнях работают хуже. Затрудняющими факторами являются малая длина единиц нижних уровней, высокая вариативность, возможность вложения единиц одного уровня друг в друга или пересечения их. Удобными полигонами для тестирования алгоритмов выделения структурных единиц нижнего уровня являются тексты на естественном языке без заглавных букв, знаков препинания и пробелов между словами. На объектах подобного рода в [6] рассматривалась задача выявления морфемной структуры слов, а в [8] – самих слов. Важно подчеркнуть, что обучающий материал отсутствовал.

Промежуточное положение между алгоритмами выделения единиц нижнего и верхнего уровня заполняют работы по построению иерархических грамматик [9] и сложностных разложений [10]. В них сегментация последовательностей осуществляется с привлечением единиц разных иерархических уровней. Оба подходы успешно используются для сжатия данных.

2 Возможные подходы к формализации понятия "структурная единица"

2.1 Сложностной анализ

В основе большинства методов выделения структурных единиц нижнего уровня лежит понятие *повтора в широком смысле*. Различают повторы прямые и симметричные, следующие друг за другом (тандемные) и разнесённые, совершенные (точные) и несовершенные (с искажениями). Возможны повторы с переименованиями элементов алфавита или повторы с точностью до фиксированного агрегирования элементов алфавита. При отсутствии дополнительной априорной информации *аномально* длинные или частые, или редкие повторы разных типов уже могут рассматриваться как потенциально возможные структурные единицы. Аномальность определяется в сопоставлении со значениями, ожидаемыми для указанных параметров в случайных последовательностях той же длины и с тем же частотным составом элементов.

Если привлечь позиционную информацию, т.е. информацию о местах вхождения повторов в текст, то интерес будут представлять участки текста с аномально высокой концентрацией повторов разного типа. Объективным количественным индикатором насыщенности участка текста повторами может служить формально вычисляемый показатель сложности. Универсальная (в смысле применимости к различным языковым системам) мера сложности конечной символьной последовательности была предложена Лемпелем и Зивом [10]. В рамках их подхода сложность последовательности оценивается числом шагов порождающего её процесса. Допустимыми (редакционными) операциями при этом являются: а) генерация

символа (необходима, как минимум, для синтеза элементов алфавита) и б) копирование "готового" фрагмента из предыстории (т.е. из уже синтезированной части текста).

Пусть Σ – конечный алфавит, S – текст (последовательность символов), составленный из элементов Σ; $S[i]$ – i-й символ текста; $S\ [i : j]$ – фрагмент текста с i-го по j-й символ включительно ($i < j$); $N = |S|$ – длина текста S. Тогда схему синтеза последовательности S можно представить в виде конкатенации

$$H(S) = S[1 : i_1]S[i_1 + 1 : i_2] \dots S[i_{k-1} + 1 : i_k] \dots S[i_{m-1} + 1 : N], \quad \text{.....(1)}$$

где $S[i_{k-1} + 1 : i_k]$ – фрагмент S, порождаемый на k-м шаге, а $m = m_H(S)$ – число шагов процесса. Из всевозможных схем порождения S выбирается минимальная по числу шагов. Таким образом, сложность последовательности S по Лемпелю и Зиву

$$c_{LZ}(S) = \min_H \{m_H(S)\}\ .$$

Минимальность числа шагов обеспечивается выбором для копирования на каждом шаге максимально длинного прототипа из предыстории. Если обозначить через $j(k)$ номер позиции, с которой начинается копирование на k-м шаге, то длина копируемого фрагмента

$$l_{j(k)} = i_k - i_{k-1} - 1 = \max_{j \le i_{k-1}} \{l_j : S[i_{k-1} + 1\ : i_{k-1} + l_j] = S[j : j + l_j - 1]\}, \qquad (2)$$

а сам k-й компонент сложностного разложения (1) можно записать в виде[1] :

$$S[i_{k-1} + 1 : i_k] = \begin{cases} S[j(k) : \quad j(k) + l_{j(k)} - 1] & \text{при} \quad j(k) \neq 0, \\ S[i_{k-1} + 1] & \text{при} \quad j(k) = 0. \end{cases} \qquad (3)$$

Случай $j(k) = 0$ соответствует ситуации, когда в позиции $i_{k-1} + 1$ стоит ранее не встречавшийся символ и мы применяем операцию генерации символа.

Пример 1. Пусть Σ = {A,B} и S = ABBABAABBAABABBA. Схема порождения S имеет вид:

$H(S)$ =	A ·	B ·	B ·	AB ·	A ·	ABBA ·	ABA ·	BBA;	$c_{LZ}(S) = 8$
$j(k)$:	0	0	2	1	4	1	4	8	

Здесь компоненты разложения отделены друг от друга точками. Само разложение можно трактовать как представление текста в терминах повторов, среди которых присутствуют наиболее значимые (самые длинные).

Сложностной анализ текста можно проводить в двух режимах – *сегментации* и *фрагментации*. Первый режим рассмотрен выше. Он даёт интегральное представление о структуре последовательности в целом и сводится к разбиению её на непересекающиеся, но взаимосвязанные сегменты (без пробелов). Другой режим сводится к поиску отдельных фрагментов, характеризующихся *аномально низкой сложностью*, т.е. достаточно *высокой степенью структурированности*. Такие фрагменты выявляются с помощью вычисления локальной сложности в пределах окон переменной длины, скользящих вдоль последовательности. Кривые изменения локальной сложности вдоль последовательности называются *сложностными профилями*. Набор профилей при разных размерах окон позволяет выявить границы аномальных фрагментов и их взаимосвязи. Примеры выявления функционально значимых фрагментов в биологических последовательностях с помощью ДНК-ориентированной меры сложности будут приведены ниже (детали см. в [11]).

В текстах на естественном языке аналогом ДНК-фрагментов с аномально низкой сложностью являются "сверхфразовые единства" – достаточно крупные фрагменты, соотносимые с отдельными микротемами текста. Связующими элементами в них выступают

---

[1] Это несколько упрощенный вариант меры LZ, допускающий наличие одинаковых компонентов в разложении (1). В исходной работе [10] все компоненты разложения уникальны, что достигается приписыванием на каждом шаге к копируемому фрагменту ещё одного (очередного) символа путем использования операции генерации.

ассоциативно связанные друг с другом кластеризованные знаменательные словоформы. Для выявления позиционной кластеризации удобно использовать сканирующие статистики. Примером может служить статистика $d(n)$, равная длине минимального фрагмента текста, содержащего ровно n вхождений нормализованной словоформы x ($n_{\text{пор}} \le n \le f(x)$), где $f(x)$ – частота встречаемости словоформы x в тексте, а $n_{\text{пор}}$ – ограничение снизу на число её вхождений в кластер. Кластер может быть зафиксирован при любом значении n, если значение $d(n)$ аномально мало по сравнению с ожидаемым при случайном распределении словоформы x по длине текста.

Участки кластеризации разных словоформ могут быть вложены один в другой, разнесены по тексту или пересекаться друг с другом. Совокупное распределение по длине текста и взаимосвязь кластеризованных структурных единиц (слов и словосочетаний) отражает *профиль кластеризуемости* лексических единиц [12], который можно рассматривать как аналог сложностного профиля. Формально, профиль кластеризуемости – это ступенчатая функция, аргументом которой является порядковый номер предложения в тексте, а значение равно числу различных кластеров, включающих в себя данное предложение. Профиль кластеризуемости хорошо отражает макроструктуру текста

2.2 Критерии устойчивости

Для выделения самых мелких структурных единиц, подобных морфемам и словам естественного языка, в той или иной форме используется понятие "устойчивости". Наряду с повторяемостью его можно рассматривать как элемент определения структурной единицы, закрепляющий её самостоятельный статус. Термин впервые был употреблён, по-видимому, в работе [6]. В основе оценки устойчивости лежат следующие наблюдения: 1) предъявление части структурной единицы в значительной мере обеспечивает возможность прогноза её оставшейся части; 2) в тексте структурная единица функционирует в разнообразных окружениях. Первое свойство в [6] названо внутренней устойчивостью – St_{int}, второе – внешней устойчивостью – St_{ext}. Иными словами, чем длиннее предъявленная цепочка символов, начинающих (или заканчивающих) структурную единицу, тем с большей вероятностью можно угадать её окончание (или начало). На границах же структурной единицы степень неопределённости предсказания возрастает. Например, слово "чемпион" мы можем легко угадать уже по первым четырём буквам, тогда как продолжение его угадать труднее: это может быть "чемпионка", "чемпионский", "чемпионат" и т.п.

Пусть u будет произвольная цепочка символов длины d из текста S, $f(u)$ – частота её встречаемости в S. При всевозможных разбиениях цепочки u на две части её можно представить в виде конкатенации левой и правой частей: $u = l_i r_i$, где $|l_i| + |r_i| = d$, а индекс i характеризует конкретное разбиение ($1 \le i \le d - 1$). Тогда успешность прогнозирования правой части цепочки по левой можно оценить отношением $f(u) / f(l_i)$, а левой части по правой – $f(u) / f(r_i)$. При хорошем прогнозировании эти величины близки к 1, при плохом близки к нулю. Внутренняя устойчивость цепочки u оценивается как средняя её прогнозируемость по обеим частям при всевозможных разбиениях u на две части:

$$St_{\text{int}}(u) = \frac{1}{2(d-1)} \sum_{i=1}^{d-1} \left(\frac{f(u)}{f(l_i)} + \frac{f(u)}{f(r_i)} \right) \tag{4}$$

Для цепочек длины 1 St_{int} полагается равной нулю. Внешняя устойчивость в [6] оценивается всего двумя градациями:

$$St_{ext}(u) = \begin{cases} 0, & \text{если } u \text{ вложена в } v \quad \text{и } f(u) = f(v), \\ 1, & \text{в противном случае} \end{cases} \tag{5}$$

Нулевое значение St_{ext} означает, что цепочка u не имеет самостоятельного статуса, поскольку в тексте S она функционирует лишь в составе более длинной цепочки v. Поскольку частоты достаточно длинных цепочек обычно равны 1[2] , равно как и частоты их расширений, внешняя устойчивость таких цепочек равна 0.

[2] Это выполняется для всех цепочек, длина которых превышает длину максимального повтора в тексте.

Полная устойчивость цепочки u определяется как произведение обеих устойчивостей:

$$St(u) = St_{\text{int}}(u) \cdot St_{ext}(u)$$

Определение внутренней устойчивости представляется не слишком удачным, поскольку хорошее прогнозирование имеет место лишь, когда предъявлена значительная часть структурной единицы (начальная или конечная). Попытки предсказания на основе коротких цепочек (1 – 2 буквы) создают ненужный "шумовой фон" в выражении (4). Оценка внешней устойчивости слишком грубая. Если значения $f(u)$ и $f(v)$ в (5) не слишком малы, то небольшие различия между ними следует скорее трактовать как "неустойчивость" ($St_{ext} = 0$), чем наоборот.

В [8] та же по сути характеристика, что и устойчивость, определяется в терминах шенноновской энтропии. Структурная единица должна характеризоваться относительно низкой энтропией внутри (хорошая предсказуемость целого по части) и относительно высокой энтропией на границах (большая неопределённость в предсказании элемента, расположенного непосредственно слева или справа от структурной единицы).

Оценка внешней устойчивости цепочки u через энтропию выглядит удачнее, чем в [6]. Пусть $ua_1, ua_2 \ldots, ua_m$ будут всевозможные правосторонние[3] расширения цепочки u на один символ a_i ($a_i \in \Sigma$), зафиксированные в тексте ($1 \le i \le m$, m – число различных расширений). Тогда энтропийная версия внешней устойчивости выглядит следующим образом:

$$H_{ext}(u) = -\sum_{i=1}^{m} \frac{f(ua_i)}{f(u)} \log \frac{f(ua_i)}{f(u)} \tag{6}$$

Очевидно, что при $m = 1$ (все вхождения u в текст имеют одно и то же продолжение) $H_{ext}(u) = 0$, что согласуется с (5). Однако при наличии разных вариантов продолжения u $H_{ext}(u)$ даёт более гибкую оценку внешней устойчивости, чем (5).

Оценка внутренней устойчивости в [8] имеет вид:

$$H_{\text{int}}(u) = \max_i (\hat{f}(l_i) + \hat{f}(r_i)), \quad 1 \le i \le d-1, \tag{7}$$

где l_i и r_i имеют тот же смысл, что и в (4), а $\hat{f}(\circ) = \dfrac{f(\circ) - \bar{f}}{s}$ – нормированные частоты. Нормировка осуществляется вычитанием из наблюдаемой частоты конкретной цепочки $f(\circ)$ средней частоты $\bar{f}$ всех цепочек той же длины, представленных в тексте, и делением разности на среднеквадратичное отклонение s. Следует однако отметить, что в отличие от внешней устойчивости (6), попытка трактовать и внутреннюю устойчивость в энтропийных терминах выглядит у авторов [8] не слишком убедительно.

2.3 Учет факультативных признаков

Рассматриваемые выше достаточно универсальные подходы к формализации понятия "структурная единица" могут быть дополнены рядом факультативных признаков, многие из которых носят общеязыковой характер. Укажем некоторые из них:

– *состав начальных и конечных элементов структурных единиц далеко не произволен*. Например, слово не может начинаться с мягкого знака, но часто им заканчивается. Кодирующие последовательности генов начинаются с инициирующих кодонов (ATG или GTG), а заканчиваются стоп-кодонами (TAA, TAG, TGA). Характерные биграммные комбинации встречаются на стыках экзонов и интронов, образующих мозаичную структуру эукариотических генов. Тематически однородные и локально завершённые фрагменты в текстах научных статей часто заканчиваются характерными индикаторами типа "подводя итог", "завершая обсуждение" и т.п., а начинаются маркерами "в данном разделе", "переходя к" и пр. Попевочные структуры в знаменных песнопениях часто начинаются переходным знаменем "голубчик борзый", а заканчиваются знаменами из семейства "статей" или "крыжом". Сложность использования таких признаков в том, что они могут встречаться и в

[3] Авторы [8] ориентируются на тексты, читаемые слева направо, и используют лишь правостороннюю оценку устойчивости.

середине структурной единицы. Если бы не это обстоятельство, их можно было бы трактовать как *формальные разделители*, характеризующиеся "*сверхравномерным*" (по отношению к случайному) распределением по тексту. Иными словами, они не должны допускать слишком большого сближения друг с другом, равно как и удаления друг от друга. Такие цепочки символов могут быть обнаружены в тексте с помощью специфических сканирующих статистик [13];

– значительный интерес в плане выявления потенциально возможных границ структурной единицы и установления взаимосвязей между единицами соседних иерархических уровней представляет *исследование вариативности структурных единиц*[4]. Общая тенденция такова: множественные искажения в виде одиночных замен, вставок, устранений символов распределены по длине структурной единицы (или повтора) не равномерно, а демонстрируют *сильную позиционную кластеризуемость*. Искажения приходятся либо на границу между единицами более низкого иерархического уровня, либо (протяжённый кластер) целиком модифицируют единицу низшего уровня [14]. Эта закономерность наблюдается не только для естественного языка, но и на музыкальных и генетических текстах;

– многие тексты бывают представлены в виде взаимосвязанных билингв, синхронизированных друг с другом (нуклеотидные и аминокислотные последовательности, стихотворные тексты и соответствующие им песенные мелодии и т.п.). Наличие информации о границах структурных единиц в одной из последовательностей (например, в стихотворном тексте) может облегчить сегментацию связанного с ним текста (в данном случае – музыкального);

– *аномально длинные повторы*, которых бывает довольно много в любом "неслучайном" тексте, как правило, являются самостоятельными структурными единицами достаточно высокого уровня. Сопоставление их друг с другом позволяет обнаружить общие фрагменты, которые можно трактовать как единицы более низкого уровня. Эта тактика срабатывает, например, при выделении попевочных структур в знаменных песнопениях [7] или поиске условно синонимичных подстановок в терминологической лексике.

3 Алгоритмические аспекты

Все описанные в предыдущем разделе возможные подходы к формализации понятия "структурная единица" и способам выделения структурных единиц из текста опираются на информацию о частоте встречаемости цепочек символов (или более крупных единиц) в тексте с привлечением (при необходимости) и позиционной информации. Необходимые данные могут быть получены на основе системы *L-граммного представления текстов*.

Термин "*L*-грамма" был впервые использован К. Шенноном применительно к цепочке x_L из L подряд следующих букв текста (L = 1,2,…). В настоящее время его используют (хотя и не совсем корректно) и по отношению к цепочкам слов. Совокупность L-грамм, описывающих текст, формируется путём анализа содержимого окна размера L (символов или слов), скользящего вдоль текста со сдвигом на одну позицию. Под *L-граммной характеристикой текста S* мы понимаем совокупность $\Phi_L(S)$ всевозможных представленных в нём L-грамм с указанием их частот встречаемости $f(x_L)$ а, при необходимости, и мест вхождения каждой из них в текст. Параметр L обычно пробегает значения от 1 до $L_{\max}(S)$, где $L_{\max}(S)$ – длина максимальной *повторяющейся* цепочки в тексте. Характеристики более высокого порядка уже практически не несут новой информации о тексте, поэтому совокупность $\Phi(S) = \{\Phi_1(S), \Phi_2(S), \ldots, \Phi_{L_{\max}}(S)\}$ мы отождествляем с полным L-граммным спектром текста.

Для вычисления L-граммных спектров могут быть использованы суффиксные деревья, L-граммные деревья (trie-структуры) или рекуррентное хеширование [15]. Суффиксное дерево содержит полную информацию обо всех L-граммах текста и их частотах и требует линейных (в зависимости от длины текста N) затрат на этапе построения. Однако для извлечения из него конкретной L-граммной характеристики $\Phi_L(S)$ требуются дополнительные затраты. В двух других методах спектр вычисляется последовательно по L ($\Phi_1(S)$, $\Phi_2(S)$ и т.д.), причем при

[4] Если таковые не выделены, можно ограничиться анализом несовершенных (содержащих искажения) повторов в тексте.

построении $\Phi_{L+1}(S)$ используется информация от предыдущей (L-й итерации). Трудоёмкость в наихудшем случае – $O(L_{\max}N)$.

К достоинствам L-граммных спектров можно отнести: возможность использования их для выделения структурных единиц разных уровней (применительно к естественному языку – от морфем до устойчивых словосочетаний и выше); невысокую трудоёмкость вычислений; возможность выявления взаимосвязей между единицами одного и того же и разных уровней; наличие обобщений на случай группы текстов. Получаемая при этом информация может быть использована для многоплановой классификации текстов [16].

Задачи сегментации, формулируемые как получение разбиения текста с максимальной суммарной устойчивостью составляющих его компонентов (см., например, [6]), достаточно сложны и для их решения используются приближенные алгоритмы. На данном этапе это оправдано, поскольку построение формального описания структурных единиц нельзя считать завершённым. Другие варианты сегментаций, возникающие при сложностном анализе текста [10] или при построении иерархических грамматик [9], просты в вычислительном отношении (линейные и квазилинейные алгоритмы), но содержат элементы разных иерархических уровней. При этом строящиеся формальные иерархии не всегда согласуются с естественными иерархиями типа "морфемы – слова – словосочетания", что особенно заметно на элементах нижних уровней.

Алгоритмы фрагментации, осуществляющие поиск наиболее "значимых" участков текста, почти все работают в режиме "скользящего окна". Это, в сочетании с рекуррентным пересчётом интересующей нас характеристики (например, сложности) в скользящем окне, гарантирует их невысокую трудоёмкость.

4 Примеры анализа реальных текстов. Учёт специфики конкретных языковых систем

Изложенные в разделе 1 довольно общие принципы, закладываемые в определение структурных единиц, претерпевают довольно существенную коррекцию при учёте специфики конкретных языковых систем. Приведём некоторые примеры на эту тему.

4.1 Сложностные профили ДНК-последовательностей

Характерной особенностью ДНК-последовательностей является широкий спектр представленных в них повторов. Многие из них обусловлены отношением комплементарности, в соответствии с которым попарно связаны нуклеотиды в двойной спирали (A с T, C с G). В соответствии с этим будем различать повторы четырёх типов:

а) прямые …AGCTTA…AGCTTA… (повторы в обычном смысле; будем выделять их подчеркиванием);

б) симметричные: ...$\overleftarrow{\text{AGCTTA}}$...$\overrightarrow{\text{ATTCGA}}$.. (выделяются расходящимися стрелками сверху);

в) прямые комплементарные: …$\overrightarrow{\text{AGCTTA}}$…$\overrightarrow{\text{TCGAAT}}$… (прямые повторы с точностью до переименования элементов алфавита: A→T, T→A, C→G, G→C; выделяются одинаково направленными стрелками сверху);

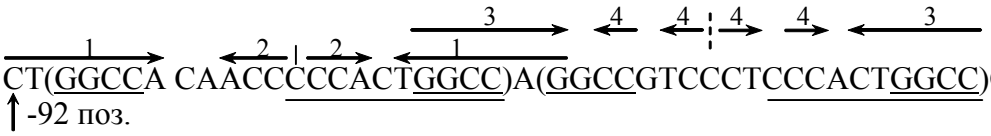
г) симметричные комплементарные: …$\overrightarrow{\text{AGCTTA}}$…$\overleftarrow{\text{TAAGCT}}$… (симметричные повторы с точностью до подстановки A→T, T→A, C→G, G→C; выделяются сходящимися стрелками сверху).

Значимость и функциональная нагрузка повторов типа "а" и "г" не вызывает сомнений. Вопрос же о существовании неслучайных повторов типа "б" и "в", механизмах их возникновения и функциональной нагрузке дебатируется.

Операция копирования, фигурирующая в определении меры сложности Лемпеля и Зива, фиксирует повторы в традиционном понимании (типа "а"). В этом плане она универсальна, т.е. применима к текстам любой языковой природы. Если мы хотим при анализе нуклеотидных последовательностей учесть все четыре типа повторов, мера становится ДНК-

ориентированной[5]. Смысл предложенной нами модификации состоит в том, чтобы вместо одной ввести четыре операции копирования, соответствующие указанным выше типам повторов, и использовать на каждом шаге моделирования последовательности ту из операций, которая максимальным образом удлиняет уже синтезированный фрагмент. Нижеследующий пример демонстрирует суперпозицию многих структур, выделенных с помощью сложностного анализа во фрагменте промотора[6] из подборки (Bucher).

Пример 2. Аномально низкий по сложности фрагмент промотора HSLCATG. Указана лишь часть структур.



Здесь фрагменты в круглых скобках образуют длинный несовершенный повтор, фланкированный короткими повторами CT (консервативные участки, в частности GGCC, подчеркнуты). 4-кратное повторение комплементарного палиндрома GGCC приводит к образованию структур шпилечного типа (1) и (3) с элементами симметрии (2) и (4) в петлевых участках.

С помощью сложностного профиля можно проследить и структуру генома в целом. На рис. 1 приведен график сложностного профиля генома вируса Эпштейна-Барр.

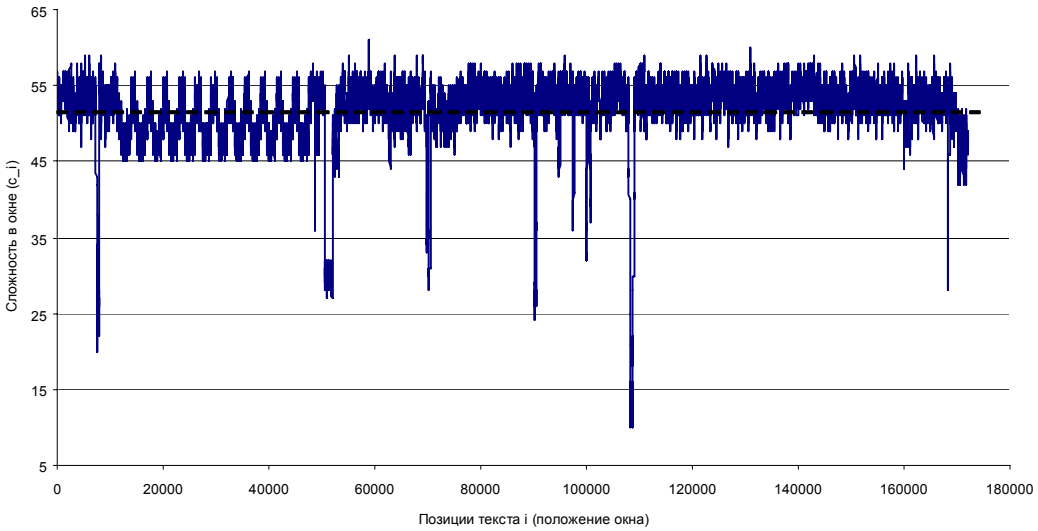


Рис. 1. Профиль сложности последовательности "Epstein-Barr virus (EBV)".

Длина текста – 172281, размер окна – 200.

Среднее значение сложности в окне – 52.85 (обозначено пунктиром), s = 4,77.

Главный минимум (ему соответствует значение сложности в окне, равное 9) обусловлен не менее чем 30-кратным варьированным повторением цепочки из 18 нуклеотидов в кодирующей области (EBNA-1 протеин). Очень интересно поведение профиля в районе поз. 12000 – 50000. Он не содержит аномально низких значений, но демонстрирует строгую периодичность. Объясняется это тем, что в данном районе имеет место 12-кратное повторение фрагмента из 3072 нуклеотидов, однако малый размер окна анализа (200 нк) не позволяет отреагировать на это снижением сложности. Если сделать размер окна больше, чем 3072 (например, 5000 нк) главный минимум переместится в эту область.

---

[5] Заметим, что в общем случае мы можем обнаруживать в тексте повторы с точностью до произвольной (и заранее не заданной!) подстановки на элементах алфавита (их всего $n!$, где n – размер алфавита). Мера вновь становится универсальной, при этом трудоёмкость алгоритма возрастает несущественно (см. [11]).

[6] Промотор – регуляторная часть гена, с которой начинается транскрипция кодирующей части.

4.2 Выделение устойчивых цепочек слов

Специфика данной задачи состоит в том, что исходный текст уже структурирован (выделены слова, предложения и т.п.). Поэтому единицами нижнего уровня для нас уже являются слова, а не символы. Далее, мы должны уметь отождествлять варианты одной и той же канонической формы с учётом её склонения, спряжения и т.п. Поэтому все словоформы текста предварительно нормализуются с использованием процедуры морфологического анализа. И, наконец, формирование цепочек слов происходит лишь в рамках одного предложения, поскольку выделяются единицы уровня, промежуточного между словом и предложением.

Критерий устойчивости используем в форме, упрощённой по сравнению с той, что описана в разделе 2.2. Если при расширении цепочки влево или вправо существует доминирующее продолжение (например, такое, частота которого превышает сумму частот всех остальных вариантов), считаем, что формирование структурной единицы не закончено (цепочка неустойчива, процесс расширения продолжается). В противном случае фиксируем границу структурной единицы (правую или левую).

Формально, пусть x_L – произвольная L-грамма, составленная из слов, $f(x_L)$– частота её встречаемости в тексте. Из всевозможных (реализованных в тексте) левосторонних расширений цепочки x_L, имеющих форму ax_L, где a – произвольная словоформа, предшествующая x_L, выберем расширение a^*x_L с максимальной частотой встречаемости – $f(a^*x_L)$. Очевидно, что $f(a^*x_L) \le f(x_L)$. Аналогично, среди всех правосторонних расширений x_Lb выберем самое частое – x_Lb^*, для которого, в свою очередь, справедливо соотношение $f(x_Lb^*) \le f(x_L)$. Цепочка x_L с $f(x_L) \ge 2$ считается устойчивой, если одновременно выполняются соотношения: $f(a^*x_L) / f(x_L) \le \Pi$ и $f(x_Lb^*) / f(x_L) \le \Pi$. Пороговое значение Π можно выбрать в районе 0,5. Эти неравенства исключают доминирование одного из возможных расширений.

Полезным дополнением к описанной методике является анализ вариативности выделенных устойчивых цепочек. Он предполагает наличие в тексте цепочек, отличающихся от конкретной выделенной либо условно-синонимичной заменой в одной из позиций ("в настоящей работе", "в данной работе" и т.п.), либо вставкой ограниченной длины ("в настоящей работе", "в настоящей и предшествующей работах"). Они могут быть не выделены как самостоятельные устойчивые цепочки, например, из-за низкой частоты встречаемости (однократное вхождение). Такого рода цепочки, составляющие ближайшую окрестность выделенной устойчивой цепочки, могут быть легко обнаружены с помощью анализа L-граммных характеристик того же (в случае замен) или более высокого (в случае вставок) порядка (см. [14]).

В заключение данного раздела отметим, что аналогичный подход был успешно использован для выделения устойчивых цепочек знамен, отождествляемых с попевками – элементарными структурными единицами знаменного распева [7]. Спецификой этой (очень интересной) языковой системы является наличие билингв типа "знамя - нота", насыщенность тандемными повторами с разной нотолинейной интерпретацией, характерные комбинации символов в конце, начале, а иногда и в середине структурной единицы, проявления "тайнозамкненности" (аналогом могут служить фразеологизмы в естественном языке) и др.

5 Заключение

Выделение и интерпретация структурных единиц в символьных последовательностях различной языковой природы – важный этап в автоматизации построения онтологий для многих предметных областей. Рассмотрены возможные подходы к формализации понятия "структурная единица". Наряду с общеязыковыми закономерностями отмечены специфические, характерные для конкретных языковых систем. Многие подходы доведены до алгоритмов и апробированы на реальных данных. Приведены примеры использования этих подходов в разных языковых системах (обнаружение регуляторных структур в ДНК-последовательностях, выделение терминологических словосочетаний в текстах предметной области, уточнение номенклатуры структурных единиц знаменного распева и формирование словаря попевок, составляющих его основу). Проблемными остаются вопросы формализации отдельных свойств структурных единиц при отсутствии обучающих подборок, учёта вариативности, интеграции различных формальных подходов в единый программный комплекс, а также интерпретации формально выделяемых структурных единиц.

Литература

[1] Jerom V. Braun and Hans-Jeorg Müller: Statistical Methods for DNA Sequence Segmentation. *Statistical Science*, 1998. Vol.13, No. 2, P. 142-162.

[2] Fu, Y.-X. and Curnow, R.N.: Maximum likelihood estimation of multiple change point. *Biometrica*, 1990, Vol. 77, P. 563-573.

[3] Hartigan J.A.: Partition models. *Comm. Statist. Theory Methods*, 1990, Vol. 19, P.2745-2756.

[4] Churchill J.A.: Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 1989, Vol. 51, P. 79-94.

[5] Lin J.: Divergence measures based on the Shannon entropy. *IEEE Trans. on Information Theory*, 1991, Vol. 37, P. 145-151.

[6] Сухотин Б.В.: Оптимизационные методы исследования языка. М., Наука, 1976.

[7] Бахмутова И.В., Гусев В.Д., Титкова Т.Н.: *L*-граммные азбуки для дешифровки знаменных песнопений. *Сибирский журнал прикладной и индустриальной математики*, 1998, Т. 1, № 2, С. 51-66.

[8] Paul Cohen, Niall Adams and Brent Heeringa: Voting experts: An unsupervised algorithm for segmenting sequences. *Intelligent Data Analysis*, 2007, Vol. 11, P. 607-625.

[9] Craig G. Nevill-Manning, Ian H. Witten.: Identifying Hierarchical Structure in Sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, 1997, Vol. 7. P. 67-82.

[10] Lempel A., Ziv J.: On the complexity of finite sequences. *IEEE Trans. on Information Theory,* 1976, Vol. IT-22, No. 1. P. 75-81.

[11] Vladimir D. Gusev, Lubov A. Nemytikova and Nadia A. Chuzhanova: On the complexity measures of genetic sequences. *Bioinformatics*, Vol.15, No.12, 1999, 994–999.

[12] Гусев В.Д., Мирошниченко Л.А., Саломатина Н.В.: Тематический анализ и квазиреферирование текста с использованием сканирующих статистик . *Труды международной конф. Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии"*, Звенигород, 1-7 июня 2005. - М., Наука, 2005, С. 121–125.

[13] Гусев В.Д., Немытикова Л.А., Саломатина Н.В.: Выявление аномалий в распределении слов или связных цепочек символов по длине текста. *Интеллектуальный анализ данных* (Вычислительные системы, вып. 171), Новосибирск, 2002, С. 51—74.

[14] Саломатина Н.В.: Методы и программные средства выделения и численного оценивания вариативности языковых единиц. *Автореферат диссертации на соискание ученой степени кандидата физико-математических наук по специальности* 05.13.11 – "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей", 2009.

[15] Гусев В.Д., Немытикова Л.А.: Учет проявлений повторности, симметрии и изоморфизма в символьных последовательностях. *Методы обнаружения эмпирических закономерностей* (Вычислительные системы, вып. 167), Новосибирск, 2001, С. 11 – 33.

[16] Гусев В.Д., Саломатина Н.В.: *L*-граммное представление текстов на естественном языке и его возможности. *Материалы Всероссийской научной конференции Квантитативная лингвистика: исследования и модели* (КЛИМ–2005), Новосибирск, 6–10 июня 2005, 256–270.