УДК 510.62

А. С. ЛЕВИЦКИЙ, Н. В. ШАРОНОВА, канд. техн. наук,
Э. М. БУЗНИЦКАЯ, канд. техн. наук

ИСПОЛЬЗОВАНИЕ ЛИНГВИСТИЧЕСКОГО РЕГИСТРА ПРИ РЕШЕНИИ ЗАДАЧ АНАЛИЗА И СИНТЕЗА РУССКОГО ТЕКСТА

При построении и эксплуатации различных информационных систем предпочтительна естественно-языковая форма общения пользователя и ЭВМ. Ведутся активные работы по моделированию естественного языка и его фрагментов. В процессе обработки текстовой информации важнейшими являются задачи анализа и синтеза словоформ, включающие морфологическую и словообразовательную (деривативную) обработку.

Рассмотрим русское суффиксальное словообразование на примере отглагольных имен существительных. В работах [1, 2] было описано построение модели смысла, а в данной статье основное внимание уделяется построению модели текста, описанию его структуры и связей текста суффиксальной морфемы с моделью смысла.

Введем некоторую вспомогательную абстрактную конструкцию, называемую лингвистическим регистром (ЛР). Это обусловлено требованиями, предъявляемыми к тексту как к объекту описания математическим аппаратом алгебры конечных предикатов (АКП). Для выбора и введения переменных, связывающих уравнениями фрагменты текста, необходимо каждому из них (букве, суффиксальному сегменту, морфу и т. д.) поставить в соответствие некоторые переменные, обозначающие место данного фрагмента в тексте. Эти функции выполняет ЛР. Кроме того, он является удобной промежуточной конструкцией для представления наиболее экономным образом текстовых единиц в памяти ЭВМ.

Лингвистический регистр сегментированных суффиксов (ЛРСС) [3] — удобное вспомогательное средство для построения математических моделей отношений, связывающих суффиксальную часть словоформы с остальным текстом. Распределение букв по ЛРСС осуществляется специальным алгоритмом сегментирования текста, в результате работы которого за каждым фрагментом закрепляется определенное место в ЛРСС. Суффиксальный сегмент в ЛРСС имеет структуру <гласная буква> <согласная буква> <мягкий знак>, на каждом из перечисленных мест может находиться либо буква, либо пробел ⎵. Подробно структура и функции ЛРСС описаны в работе [3, с. 41—57].

Нами предпринята попытка использования ЛРСС не просто как вспомогательной абстрактной конструкции, роль которой состоит в присвоении набора мест фрагментам текста, но и в качестве отправной, базовой конструкции, несущей в себе

семантические связи морфов с окружающим их текстом. В процессе исследований осуществлен семантический переход от суффиксальной морфемы через сегмент суффикса к отдельной букве суффиксального сегмента. Возможно также решение обратной задачи: переход от буквы суффиксального сегмента, находящегося в ЛРСС, к набору морфов и морфем, куда этот сегмент может входить. В качестве промежуточной преследовалась цель минимизации ЛРСС на буквенном уровне. Она выполнима при условии, что семантика морфемы перенесена на минимальные ее фрагменты — буквы.

При минимизации ЛРСС ставится условие: каждая буква русского алфавита встречается в позиции регистра F_i только один раз, несет на себе семантику всех морфем, в которые она входит в составе данного сегмента ЛРСС. Эта общая установка позволяет подойти к решению основной поставленной задачи: осуществить аналитическую запись минимизированного ЛРСС в виде системы уравнений АКП. Данные уравнения описывают семантические связи помещенных в ЛРСС морфем, а также несут информацию о линейной сочетаемости суффиксальных сегментов и морфов.

Попытка минимизации ЛРСС на сегментном уровне описана в работе [3]. Минимизировалось количество сегментов морфем в F_i сегменте регистра. Иными словами, исключались повторяющиеся сегменты морфем в столбцах F_i ЛРСС. После минимизации максимальная глубина регистра по вертикали составляла 47 строк. Это удовлетворительно, поскольку рассматривалось свыше шестисот морфов всех частей речи. На меньшем множестве исходных морфем, например только на суффиксах отглагольных имен существительных, такая минимизация нецелесообразна, так как почти не дает экономии в записи по строкам. В результате исследований установлено, что при работе с меньшим, чем полное множество морфов, множеством сегментированных единиц, гораздо эффективнее использовать минимизацию ЛРСС на буквенном уровне. Опишем последовательность действий при такой минимизации.

Исключаем повторяющиеся в сегменте ЛРСС F_i буквы. При этом к семантике буквы, с которой начался процесс минимизации, добавляем семантику тех сегментов, из которых вычеркиваем букву. Таким образом, после минимизации любая буква любого сегмента F_i ЛРСС содержит в себе семантику всех сегментов морфем, которые были расположены ниже и из которых эта буква вычеркивалась в процессе минимизации. Каждая буква минимизированного ЛРСС описывается одним уравнением и встречается в i-м сегменте на j-м месте только один раз, в то время как при сегментной минимизации одна и та же буква на месте s_{ij} встречалась более одного раза.

В качестве примера рассмотрим фрагмент регистра:

F_1	F_2	F_3	F_4	M
$s_1^1 s_1^2 s_1^3$:	$s_2^1 s_2^2 s_2^3$:	$s_3^1 s_3^2 s_3^3$:	$s_4^1 s_4^2 s_4^3$	роли r^k
⎵ль :	⎵щ⎵ :	ик⎵ :	⎵⎵	2,4
⎵ль :	⎵н⎵ :	ик⎵ :	⎵⎵	5,6
аль :	⎵ ⎵ :	⎵⎵ :	⎵⎵	8

Здесь приняты следующие обозначения: F_i — сегмент ЛРСС; s_i^j — позиция букв в сегменте; M — множество семантических ролей r, соответствующее данному сегменту суффикса. Это определенный код, разработанный ранее для суффиксов имен существительных, мотивированных глаголами [2]. Уравнения для букв ЛРСС после минимизации будут иметь следующий вид:

$$s_1^1(а) = r^3;\ \ s_1^2(л) = s_1^3(ь) = r^2 \vee r^4 \vee r^5 \vee r^6 \vee r^8;$$

$$s_2^2(ц) = r^2 \vee r^4;\ \ s_2^2(к) = r^5 \vee r^6;\ \ s_3^1(и) = s_3^2(к) = r^2 \vee r^4 \vee r^6 \vee r^5.$$

К ним следует добавить также уравнения для пробелов:

$$s_1^1(\_) = r^2 \vee r^4 \vee r^5 \vee r^6;\ \ s_2^2(\_) = s_3^2(\_) = r^8;$$

$$s_2^1(\_) = s_2^3(\_) = s_3^3(\_) = s_4^1(\_) = s_4^2(\_) = s_4^3(\_) = r^2 \vee r^4 \vee r^5 \vee r^6 \vee r^8.$$

После минимизации на буквенном уровне ЛРСС имеет вид

F_1	F_2	F_3	F_4
ать	ель	ик⎵	акь
ол⎵	ан⎵	ан	оц⎵
ув	ущ	ос	ит
ер	оч	ыт	⎵⎵
ын	ич	ец	
иг	ёр	⎵⎵	
ях	ят		
ш	ыв		
⎵д	⎵ш		
б	м		
⎵	ж		
	с		
	г		
	⎵		

Общий вид уравнения каждой буквы минимизированного ЛРСС можно записать следующим образом:

$$s_i^j(\alpha) = \bigvee_{k=1}^{n} f_k(x_1, x_2, \ldots, x_m),$$

где s_i^j — j-я позиция i-го сегмента регистра, заполненная буквой

α из множества букв русского алфавита; $x_1, x_2, \ldots, x_m$ — переменные семантические признаки [2]; $f_k(x_1, x_2, \ldots, x_m)$ — k-я семантическая роль морфемы, содержащей букву α; n — число семантических ролей буквы α. Обозначим $f_k(x_1, x_2, \ldots, x_m) = r^k$, тогда уравнение, описывающее семантические связи буквы с окружающим ее текстом, запишем так:

$$s_i^j(\alpha) = \bigvee_{k=1}^{n} \left(\bigvee_{l=1}^{m} q_k^l \right) r^k.$$

Здесь q_k^l — множитель, характеризующий влияние основы. Он обозначает класс основы внутри семантической роли r^k, $l = \overline{1, m}$ — номер класса основ.

В качестве примера запишем уравнение для буквы «а», стоящей в первой позиции первого сегмента минимизированного ЛРСС:

$$s_1^1(a) = r^7 (q_7^4 \vee q_7^5) \vee r^6 (q_6^1 \vee q_6^3) \vee r^{22} (q_{22}^2 \vee q_{22}^3) \vee r^{37} \vee r^{14} (q_{14}^1 \vee q_{14}^2) \vee$$
$$\vee r^{15} q_{15}^3 \vee r^{25} q_{25}^3 \vee r^{28} \vee r^{43} q_{43}^2 \vee r^{44} q_{44}^3 \vee r^8 (q_8^3 \vee q_8^7 \vee q_8^8) \vee r^{59} \vee$$
$$\vee r^{57} (q_{57}^1 \vee q_{57}^3) \vee r^{35} q_{35}^4.$$

Это уравнение может иметь следующую содержательную интерпретацию. Буква «а» входит в состав морфов (можно их перечислить, однако список будет слишком длинным), имеющих семантические роли 6, 7, 22, 37 и т. д. (по верхним индексам ролей) и присоединяющихся к основам перечисленных классов внутри семантических ролей (верхний индекс переменной, обозначающей класс основы).

Для каждой буквы минимизированного ЛРС записывается свое уравнение. Оно учитывает семантические связи буквы с окружающим ее ближним и дальним текстом (имеется в виду влияние слова через ближайшие морфы и влияние более отдаленного контекста), а также определяет ее месторасположение в регистре. При решении задач анализа и синтеза единиц текста эти уравнения решают совместно с другими уравнениями, задающими линейную сочетаемость сегментов в суффиксах, суффиксов в слове, совместимость суффиксальных и других морфов. В результате машинных экспериментов установлено, что минимизация ЛРСС по буквам дает еще одно преимущество: нет необходимости проверять линейную сочетаемость сегментов суффикса, поскольку она уже заложена в минимизируемый ЛРСС.

Список литературы: 1. *Бондаренко М. Ф., Шаронова Н. В.* О математическом описании процессов словообразования//Пробл. бионики.—1981.—Вып. 27.—С. 83—88. 2. *Шаронова Н. В., Бузницкая Э. М.* О структуре системы признаков при моделировании словообразования//Пробл. бионики.—1983.—Вып. 31.—С. 12—19. 3. *Шаронова Н. В.* Математические модели суффиксального словообразования и их использование для автоматической обработки отглагольных имен существительных в текстах русского языка: Дис.... канд. техн. наук.—Х., 1984.—222 с.—Машинопись.

Поступила в редколлегию 18.10.85.