УДК 510.62

М. Ф. БОНДАРЕНКО, д-р техн. наук,
А. С. ЛЕВИЦКИЙ, О. А. ЛИХАЧЕВА

ИСПОЛЬЗОВАНИЕ ЛИНГВИСТИЧЕСКОГО РЕГИСТРА ПРИ РЕШЕНИИ ЗАДАЧИ ИДЕНТИФИКАЦИИ МОРФЕМЫ

Речь — естественное средство общения между людьми и использователи предпочли бы общаться с ЭВМ на естественном языке. Поскольку естественному языку не потребуется специально обучаться, то при эксплуатации различных диалоговых систем предпочтительна естественноязыковая форма общения человека с ЭВМ. В настоящее время ведутся активные работы по моделированию естественного языка и его фрагментов. Одной из важнейших задач в проблеме формализации естественного языка является автоматизированная обработка текстовой информации.

В статье рассматривается способ идентификации суффиксальных морфем на примере существительных со значением лица, получаемых в результате аффиксации от разных частей речи. Дело в том, что существительные, образованные суффиксацией от разных частей речи, являются удобным материалом для семантических исследований. На множестве суффиксальных существительных можно проследить, как меняется смысл лексической единицы после соединения с ней того или иного аффикса, какую семантику

привносит тот или иной морф, как один и тот же морф может иметь различные словообразовательные значения в зависимости от семантических особенностей мотивирующей единицы. Статья является продолжением прежних работ авторов [1].

Поскольку тексты суффиксальных морфов — это объект математического описания, то к ним должны предъявляться все требования, как и к любому объекту описания математическим аппаратом алгебры конечных предикатов [3]. Поэтому для выбора переменных, связывающих уравнениями фрагменты текста, необходимо каждому фрагменту текста (букве, морфу) поставить в соответствие некоторые переменные, обозначающие место данного фрагмента в тексте. Для этой цели и вводится некоторая абстрактная конструкция, называемая лингвистическим регистром (ЛР). Остановимся на структуре регистра.

Лингвистический регистр сегментированных суффиксов — это абстрактная конструкция, состоящая из 4-х сегментов. Каждый сегмент состоит из трех разрядов и имеет такую структуру: <гласная буква> <согласная буква> <мягкий знак>. На каждом из мест сегмента может находиться соответствующая буква либо пробел [2]. В соответствии с буквенной структурой сегментов в лингвистический регистр по определяемому алгоритму «загружены» суффиксальные морфы, формально объединенные в морфемы. Таким образом, получается, что каждая буква любого морфа, входящего в ту или иную морфему, имеет четкую привязку к позиции ЛР. На языке алгебры конечных предикатов [3] математическая запись объединения морфов, уже размещенных в ЛР, в морфему (в нашем примере) <тель> имеет такой вид:

$$S_{11}^{-}\,S_{12}^{т}\,S_{17}^{-}\,S_{21}^{е}\,S_{22}^{л}\,S_{23}^{ь}\,S_{3}^{-}\,S_{4}^{-} \vee S_{11}^{и}\,S_{12}^{т}\,S_{12}^{-}\,S_{21}^{е}\,S_{22}^{л}\,S_{23}^{ь}\,S_{3}^{-}\,S_{4}^{-} \sim M^{5}, \quad (1)$$

где S_{kl}^{x} — l-я позиция k-го сегмента регистра, заполненная из множества букв русского алфавита. Но в реальной ситуации, после проведения морфемных швов, мы имеем некоторые буквенные последовательности (морфы), не имеющие никакой привязки к ЛР. Это последнее обстоятельство является серьезным препятствием для решения задачи идентификации морфем. Остановимся на одном из способов формализации «узнавания» морфов.

Итак, имеем ЛР, в которой загружены морфы, формально объединенные в морфемы. Каждая морфема имеет свое имя M^i, где i — порядковый номер морфемы. Осуществим переход от морфемы, имеющей определенное имя (M^i), к морфам, затем к сементам морфов, обозначая их именем морфемы (M^i), затем к каждой букве морфа, присваивая каждой букве имя (M^i). После присвоения каждой букве морфа, входящего в морфему, имени морфемы осуществим сжатие ЛР в вертикальном направлении. Опишем алгоритм сжатия ЛР.

При вертикальном сжатии ЛРСС ставилось условие, что каждая буква русского алфавита встречается в позиции S_{kl} только один раз, при этом ей присваиваются имена всех морфем, в состав

которых она входит. Исключаем повторяющиеся в сегменте ЛР F_k буквы, дизъюнктивно собираем имена морфем. В результате сжатия любой букве любого сегмента F_k присвоены имена всех морфем M^i, в состав которых она входит. Каждая буква минимизированного ЛРСС описывается только одним уравнением и встречается в k-м сегменте на l месте только один раз, в то время как в случае сегментной минимизации [2] одна и та же буква на месте S_{kl} встречалась более одного раза. Отметим, что алгоритм минимизации ЛРСС на буквенном уровне подробно описан в [1], однако в той работе преследовались другие цели, авторов интересовал вопрос связи смысла и текста.

Нас же интересует вопрос идентификации стандартных текстов — морфов с последующей их привязкой к уже распределенным для них местам в ЛРСС.

Продемонстрируем сжатие регистра на примере двух распространенных морфем <атор> и <тель>. Для этого рассмотрим фрагмент реального ЛРСС:

$\overset{F_1}{S_{11}}$	S_{12}	S_{13}	S_{21}	$\overset{F_2}{S_{22}}$	S_{23}	S_{31}	$\overset{F_3}{S_{32}}$	S_{33}	S_{41}	$\overset{F_4}{S_{42}}$	S_{43}	M^i
—	т	—	е	Л	b	—	—	—	—	—	—	M^5
и	т	—	е	Л	b	—	—	—	—	—	—	M^5
а	т	—	0	Р		—	—	—	—	—	—	M^{17}
я	т	—	0	Р		—	—	—	—	—	—	M^{17}
—	т	—	0	Р	—	—	—	—	—	—	—	M^{17}.

Здесь приняты следующие обозначения: F_k — сегмент ЛРСС; S_{kl} — позиция буквы в сегменте; M — множество морфем; M^i — морфема, соответствующая определенному морфу. Запишем теперь уравнения для каждой буквы сегмента, предварительно проведя сжатие:

$$S_{11}^{a} = M^{17};\ S_{11}^{я} = M^{17};\ S_{12}^{т} = M^5 \vee M^{17};\ S_{21}^{е} = M^5;$$

$$S_{21}^{0} = M^{17};\ S_{22}^{л} = M^{17};\ S_{22}^{р} = M^{17};\ S_{23}^{6} = M^5.$$

После вертикального сжатия реальный регистр сегментированных суффиксов, в который авторы «загрузили» 116 суффиксальных морфов, имел следующий вид:

F_1			F_2			F_3			F_3		
S_{11}	S_{12}	S_{13}	S_{21}	S_{22}	S_{23}	S_{31}	S_{32}	S_{33}	S_{41}	S_{42}	S_{43}

ась/ась/ас_/*ик*_

ет—/ет—/ит—/о— —

ир—/ир—/юн—/я— —

ов—/ов—/ял—/— — —

ун—/ун—/ыч—/— — —
юл—/юл—/—к—/— — —
яч—/яч—/—х—/— — —
ыг—/ыц—/—щ—/— — —
—х—/—й—/— — —/— — —
—ш—/—ш—/— — —/— — —
—щ—/—щ—/— — —/— — —,

а реальные уравнения для букв, входящих в выбранные нами морфемы <атор> и <тель>, выглядит так:

$$S^{a}_{11} \sim M^{3} \vee M^{17} \vee M^{18} \vee M^{20} \vee M^{21} \vee M^{35} \vee M^{36} \vee M^{41}, \quad (2)$$

$$S^{т}_{12} \sim M^{3} \vee M^{17} \vee M^{20} \vee M^{35} \vee M^{36}, \quad (3)$$

$$S^{0}_{21} \sim M^{17} \vee M^{19} \vee M^{23}, \quad (4)$$

$$S^{р}_{22} \sim M^{9} \vee M^{17} \vee M^{19} \vee M^{23} \vee M^{37} \vee M^{42} \vee M^{43}, \quad (5)$$

$$S^{е}_{u} \sim M^{3} \vee M^{5} \vee M^{6} \vee M^{11} \vee M^{13} \vee M^{16} \vee M^{19} \vee M^{37} \vee M^{53}, \quad (6)$$

$$S^{л}_{22} \sim M^{5} \vee M^{21} \vee M^{31}, \quad (7)$$

$$S^{б}_{23} \sim M^{5} \vee M^{9} \vee M^{31} \vee M^{42}. \quad (8)$$

Уравнение для каждой буквы регистра можно записать в общем виде:

$$S^{\alpha}_{ke} \sim \overset{n}{\underset{i=1}{V}} M^{i}, \quad (9)$$

где S — l-я позиция k-го сегмента регистра, заполненная буквой α из множества букв русского алфавита; M^i — морфема M, имеющая условный номер i. Можно записать уравнения алгебры конечных предикатов [3], аналитически описывающие ЛРСС после минимизации:

$$z^{a}_{\beta} \sim F_1 \vee F_2 \vee F_3 \ (10), \quad z^{b}_{\beta} \sim F_1 \vee F_2 \ (11), \quad z^{г}_{\beta} \sim F_1 \quad (12),$$

$$z^{c}_{\beta} \sim F_1 \vee F_2 \ (13), \quad z^{u}_{\beta} \sim F_1 \vee F_2 \vee F_3 \vee F_4 \ (14), \quad z^{u}_{\beta} \sim F_2 \quad (15),$$

$$z^{k}_{\beta} \sim F_3 \vee F_4 (16), \quad z^{л}_{\beta} \sim F_1 \vee F_2 \vee F_3 \ (17), \quad z^{м}_{\beta} \sim F_1 \vee F_2 \quad (18),$$

$$z^{н}_{\beta} \sim F_1 \vee F_2 \ (19), \quad z^{x}_{\beta} \sim F_1 \vee F_3 \quad (20), \quad z^{c}_{\beta} \sim F_1 \vee F_2 \vee F_3 \quad (21),$$

$$z^{т}_{\beta} \sim F_1 \vee F_2 \vee F_3 \quad (22), \quad z^{м}_{\beta} \sim F_1 \quad (23), \quad z^{2}_{\beta} \sim F_2 \vee F_3 \quad (24),$$

$$z^{ц}_{\beta} \sim F_1 \vee \quad (25), \quad z^{z}_{\beta} \sim F_1 \vee F_2 \vee F_4 \ (26), \quad z^{ш}_{\beta} \sim F_1 \vee F_2 \quad (27),$$

$$z^{щ}_{\beta} \sim F_2 \vee F_3 \quad (28), \quad z^{ы}_{\beta} \sim F_1 \vee F_2 \vee F_3 \quad (29), \quad z^{ь}_{\beta} \sim F_1 \vee F_2 \quad (30),$$

$$z^{ю}_{\beta} \sim F_1 \vee F_2 \vee F_3 \ (31), \quad z^{я}_{\beta} \sim F_1 \vee F_2 \vee F_3 \vee F_4, \quad (32)$$

где (10) — предикат, описывающий возможное появление буквы α в одном из сегментов регистра; β — порядковый номер буквы в суффиксе, поступившем для идентификации, так, например, для морфа -тель $\beta=1$, если $\alpha=$Т; $\beta=2$, если $\alpha=e$ и т. д. Запишем структуру сегмента (см. выше) в аналитическом виде:

$$S_{k1}^{(г)} S_{k2}^{(в)} S_{k3}^{(ь)} \vee S_{k1}^{(г)} S_{k2}^{(с)} \vee S_{k2}^{(с)}, \qquad (33)$$

где запись $S_{k1}^{(г)}$ означает, что в позиции 1 сегмента k может находиться только гласная буква; $S_{k1}^{с}$ — в позиции 2 сегмента K можно встретить лишь согласную букву; $S_{k3}^{ъ}$ — в позиции 3 сегмента K встречается лишь ь (мягкий знак).

Итак, пусть на вход нашей системы пришла некоторая последовательность букв <тель>. Если это морф, то его нужно идентифицировать. Воспользовавшись предикатом (33), проведем сегментные швы

$$z_1^T \sim F_k \ (34), \quad z_2^e z_3^l z_4^b \sim F_{k+1}. \qquad (35)$$

Предикаты (34) и (35) могут иметь такую содержательную интерпретацию. Буква <$т$> представляет собой сегмент K, последовательность <ель> представляет собой сегмент $K+1$. Поскольку буква <$т$> может встречаться в сегментах F_1, F_2, F_3 (22), то для идентификации морфемы можно воспользоваться предикатом (3), относящимся к сегменту $F_1(K=1)$. Поскольку оставшиеся буквы нашей последовательности находятся в одном сегменте, следующем за сегментом $K(1)$, то воспользуемся предикатами (3), (6), (7), (8). Логически умножив левые и правые части предикатов, можно получить условный номер морфемы, в которую входит исследуемый нами морф

$$S_{12}^T S_{21}^e S_{22}^l S_{23}^b \sim M^3. \qquad (36)$$

Системой предикатов типа (2—8) можно воспользоваться для обнаружения ошибок и их коррекции. Авторам удалось 116 морфов описать системой из 53-х уравнений. По сравнению с ранее применявшимся способом описания морфов этот способ несколько экономнее. К достоинствам предлагаемого способа можно отнести и его простоту.

Список литературы: 1. *Левицкий А. С., Шаронова Н. В., Бузницкая Э. М.* Использование лингвистического регистра при решении задач анализа и синтеза русского текста//Пробл. бионики. 1987. Вып. 38. С. 16—18. 2. *Шаронова Н. В.* Математические модели суффиксального словообразования и их использование для автоматической обработки отглагольных имен существительных в текстах русского языка. Автореф. дис. ... канд. техн. наук. Х., 1984. 213 с. 3. *Шабанов-Кушнаренко Ю. П.* Теория интеллекта: Математические средства. Х., 1984. 144 с. 4. *Бондаренко М. Ф.* Математические модели морфологических и фонетических отношений и их применение для автоматизации обработки речевых сообщений. Автореф. дис. ... д-ра техн. наук. Х., 1984. 350 с.

Поступила в редколлегию 15.03.89