УДК 62.506.2

В. А. ЛОВИЦКИЙ, канд. техн. наук, *В. Я. ТЕРЗИЯН*

КОДИРОВАНИЕ СЛОВ В ТВ-СТРУКТУРЕ

Для решения задач морфологического анализа слов диалоговой естественноязыковой системой (ДЕСТА), разрабатываемой в Харьковском институте радиоэлектроники, была предло-

жена ТВ-структура [1]. Идея построения такой структуры основана на том факте, что слова ряда естественных языков состоят как бы из двух частей. Первая — вещественная имеет предметную отнесенность, вторая — формальная часть выражает грамматическое значение слова [2]. Например, слово *стекло* состоит из вещественной части *стекл* и формальной — *О*. Первая часть позволяет сравнивать это слово со словами *стекла*, *стеклам*,

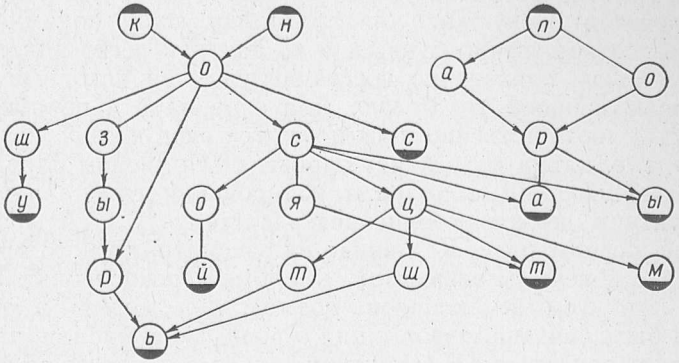

Рис. 1

стеклянный, *стекать* и т. д. Вторая говорит о грамматическом сходстве с такими словами, как *весло*, *колесо*, *бежало* и т. д. ТВ-структуру удобно представлять в виде ориентированного графа, вершины которого представляют собой буквы слов, а дуги указывают на связи этих букв в словах.

Рассмотрим принцип кодирования слов, хранящихся в ТВ-структуре. Основное требование, предъявляемое к кодам слов,— слова с одинаковой вещественной частью должны иметь одинаковые коды. Трудности, возникающие при кодировании слов, видны на конкретном примере. Пусть задано конечное множество слов *X*-{*косить*, *кошу*, *косишь*, *косит*, *косим*, *косят*, *коса*, *кос*, *косы*, *косой*, *пара*, *пора*, *поры*, *нос*, *козырь*, *корь*}. Последовательно воспринимая слова из заданного множества, ДЕСТА строит ТВ-структуру по алгоритму, приведенному в [1]. Полученная структура представлена на рис. 1. У элементов, в которые занесены начальные буквы слова, заштрихована верхняя часть кружка, у элементов с конечными буквами слова — нижняя часть кружка. Эти элементы будем называть соответственно **начальными** и **конечными** вершинами ТВ-структуры. Элемент ТВ-структуры, в котором записана буква a_i, — a_i-элемент или элемент a_i. С каждым конечным элементом связана морфологическая информация, характеризующая класс слов с одинаковой формальной частью. Ограничим пока морфологическую информацию только вопросами, на которые отвечают слова. Так, для

нашего примера каждому конечному элементу можно поставить в соответствие следующее множество вопросов: **у** — {*что делаю*}, **ь** — {*что, что делать, что делаешь*}, **й** — {*чем, какой*}, **с** — {*чего, что*}, **а** — {*что*}, **ы** — {*что, чего*}, **т** — {*что делают, что делает*}, **м** — {*что делаем*}.

Под кодом слова будем понимать минимальный набор букв, который совместно с морфологической информацией позволяет однозначно восстановить путь в ТВ-структуре, соответствующий определенному слову. Так, для слов *кошу*, *ношу* кодами будут служить соответственно буквы **к** и **н**, так как легко видеть, что по этим кодам однозначно восстанавливаются исходные слова с помощью вопроса *что делаю*. Для слов пара и пора кодами слов будут соответственно па и по, а для слов *козырь* и *корь* — кз и кр, т. е. длина кода будет определяться числом параллельных ветвей. Вводимые правила кодирования слов с помощью ТВ-структуры должны обеспечивать систему ДЕСТА морфологической информацией. Как видно из рассматриваемого примера (рис. 1), с конечным элементом ь кроме перечисленных вопросов, на которые отвечают слова *козырь*, *корь*, *косить* и *косишь*, должна быть связана информация о роде, лице, числе и падеже. Она может попасть в ТВ-структуру только в процессе ее формирования, т. е. при встрече системой ДЕСТА незнакомого слова следует вводить вместе со словом всю необходимую морфологическую информацию. Выясним, какую морфологическую информацию считать необходимой. Для настоящего уровня развития системы ДЕСТА необходимая морфологическая информация, соответствующая каждому слову, выбирается из следующего перечня: 1) вопрос, на который отвечает слово; 2) род; 3) лицо; 4) число. Все множество вопросов можно разбить на три подмножества. В первое подмножество входят вопросы, которые однозначно определяют морфологическую информацию, не заданную в явном виде. Например, вопрос *какой*? указывает на мужской род и единственное число, а *что делают*? — на третье лицо множественного числа.

Ко второму подмножеству относятся вопросы, определяющие слова, для которых морфологическая информация должна быть определена, но не непосредственно из вопроса. Например, на вопрос *что* могут отвечать слова различного рода и числа. Для слов, отвечающих на вопросы данного класса, пользователь должен в явном виде задать недостающую морфологическую информацию. Для этого достаточно ограничиться использованием одного из четырех местоимений *он*, *она*, *оно*, *они*.

Третье подмножество включает в себя вопросы к словам, которые не требуют иной морфологической информации, кроме самого вопроса. Например, когда, как и т. д.

Если работа с омонимами не вызывает затруднений (например, косой — {*чем, какой*}), то при кодировании и декодировании омографов возникают определенные трудности. Прежде все-

го слова, значения которых зависят от ударения, должны иметь различные коды. В противном случае невозможно по коду правильно восстановить слово. Например, слова *косим* (*косим траву*) и *косим* (*косим глаза*) имеют одинаковый код **к** (рис. 1) по той причине, что в ТВ-структуре не учтены ударения. В дальнейшем ударение будем обозначать апострофом, который ставится перед ударной гласной буквой. Например, *к'осим*, *кос'им*.

Таким образом, при формировании ТВ-структуры в словах должны быть указаны ударения и каждое слово должно сопровождаться морфологической информацией. После предварительного рассмотрения идеи кодирования слов с помощью ТВ-структуры перейдем к формальному описанию правил кодирования. Для этого введем ряд определений.

Для формального описания входных объектов, которые определяют вид ТВ-структуры, используем металингвистическую Бэкусову нормальную форму (БНФ). В БНФ нетерминальные символы записываются как имена, заключенные в угловые скобки <>. Символ ::= читается *заменяется на*, а символ | соответствует ИЛИ и служит для разделения альтернатив замены.

Приведем формальное описание входных объектов для русского языка. Аналогичным образом можно описать входные объекты, например, для английского языка:

<входной объект> : : = <объект 1> | <объект 2>

<объект 1> : : = <строка> <разделитель> <вопрос> <ограничитель>

<объект 2> : : = <строка> <разделитель> <местоимение> <разделитель> <вопрос> <ограничитель>

<вопрос> : : = ЧТО — ДЕЛАТЬ | ЧТО — СДЕЛАТЬ | ... | ЧТО — СДЕЛАЛИ | КТО | ЧТО | КОГО | ... | О ЧЕМ | | КАКОЙ | | ЧЕЙ | ... | КАКИЕ | ЧЬИ | КОГДА | ... | ГДЕ |

<местоимение> : : = ОН | ОНА | ОНО | ОНИ

<ограничитель> : : = Ь (где Ь — обозначает пробел)

<разделитель> : : = *

<строка> : : <русская буква> | [<строка>

<русская буква>]$_{1}^{30}$

<русская буква> : : = А | Б | В ... | Я

Для обеспечения «понимания» системой ДЕСТА все вопросы, на которые отвечают глаголы, записываются через черточку. Это позволяет избежать неоднозначности в ответе на вопрос типа: Что делает Петя? Возможные ответы: самолет или читает.

Теперь же, для получения второго ответа, вопрос должен быть записан так: Что — делает Петя?

Согласно введенному описанию преобразуем конечное множество входных объектов X следующим образом:

*X = {кос'ить * что-делать, кош'у * что-делаю, к'осишь * что-делаешь, кос'ишь * что делаешь, к'осит * что-делает, кос'ит * что-делает, кос'им * что-делаем, к'осим * что-делаем, к'осят * что-делают, кос'ят*что-делают, кос'а * она * что, кос'ы * она * чего, к'осы * они * что, к'ос * они * чего, кос'ой * какой, кос'ой * она * чем, п'ара * она * * что, п'ора * она * что, пор'а * она * что, п'оры * они *что, н'ос* он * что, к'озырь * он * что, к'орь * она * что}*

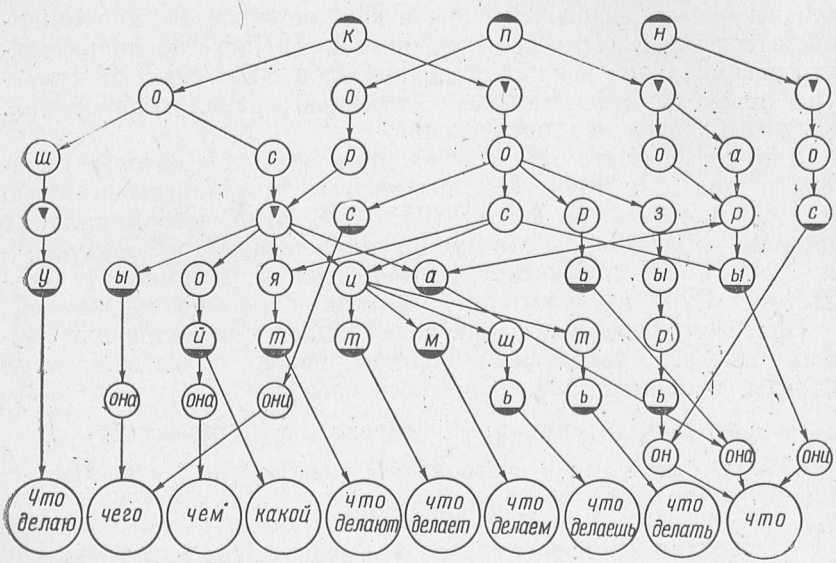

Рис. 2

Последовательно воспринимая входные объекты, системы ДЕСТА сформирует ТВ-структуру, представленную на рис. 2.

В общем случае будем считать, что входные объекты состоят из символов a_i, составляющих алфавит $\Gamma = \{a_1, a_2, \ldots, a_n\}$. Всякую конечную последовательность символов алфавита Γ будем называть **строкой** (или **цепочкой**). **Длина** строки равна числу символов в строке. Строка, которая не содержит ни одного символа, называется **пустой** строкой и обозначается символом λ.

Тот факт, что в строке x символ a_j следует за a_i, будем обозначать через $a_i \xrightarrow{x} a_j$, где символ $\rightarrow$ читается следует за, или предшествует. Непосредственное следование символа a_l за a_i

обозначим через $a_i \overset{0}{\underset{x}{\rightarrow}} a_j$, или $a_i a_j$. Если в строке $a_2 a_5 a_1 a_4 a_7$ важным является не просто сам факт следования a_7 за a_2 (в этом случае достаточно записать $a_2 \rightarrow a_7$), а то, что символ a_7 следует за a_2 через три символа, то в этом случае используется обозначение $a_2 \overset{3}{\rightarrow} a_7$. Если a_i является начальным символом строки x, а a_j — конечным, то это будем обозначать соответственно через $e \overset{0}{\underset{x}{\rightarrow}} a_i$ и $a_j \overset{0}{\underset{x}{\rightarrow}} e$, где e обозначает пустой символ.

Определение 1. Формально **строки в алфавите** Г определяются следующим образом: 1) λ строка в Г; 2) если x — строка в Г и $a_i \in$ Г, то xa_i — строка в Г; 3) x_j — строка в Г тогда и только тогда, когда она является таковой в силу 1) или 2).

Определение 2. Пусть x_i и x_j две произвольные строки. Будем говорить, что x_j **входит** в x_i, если имеет место равенство $x_i = v_1 x_j v_2$, где v_1 и v_2 — некоторые строки символов, возможно, равные λ.

Включение x_j в x_i обозначим через $x_j \subset x_i$; где символ $\subset$ читается как *входит в*.

Определение 3. Множество элементов ТВ-структуры, к которым на графе имеются пути от a_i элемента, называется **субмножеством** a_i элемента и обозначается как $\mathrm{SBS}a_i$.

Определение 4. Множество элементов ТВ-структуры, от которых на графе имеются пути к a_i элементу, называется **супермножеством** a_i элемента и обозначается как $\mathrm{SPS}a_i$.

Определение 5. Элементы из $\mathrm{SBS}a_i$, связанные с a_i элементом непосредственно, образуют **0 — субмножество элемента** a_i, или $\mathrm{SBS}^0_{a_i}$.

Определение 6. Элементы из $\mathrm{SPS}a_i$, связанные с a_i элементом непосредственно, образуют **0 — супермножество** элемента a_i, или $\mathrm{SPS}^0 a_i$.

Правило кодирования строки символов с помощью ТВ-структуры вводится следующим определением.

Определение 7. Кодом строки символов x будет служить также строка символов, которая составляется согласно правилам:

1) первым символом кода строки символов x является такой символ a_i, что

$$\left((a_i \subset x) \,\&\, \left(e \overset{0}{\underset{x}{\rightarrow}} a_i\right)\right) \vdash (\mathrm{SPS}^0 a_i = \varnothing),$$

где & — обозначает операцию конъюнкций, а $\vdash$ читается как *дает* или *приводит к*;

2) в качестве последующих символов кода последовательно берутся такие символы a_j, что

$$\exists a_i \subset x \left(\left(a_i \overset{0}{\underset{x}{\rightarrow}} a_j \right) \& \left(\left| \mathrm{SBS}^0 a_i \right| > 1 \right) \&,$$

$$\& \exists a_\nu \in (\mathrm{SBS}^0 a_i - a_j) \exists a_\mu \in \mathrm{SBS} a_\nu (a_\mu \subset x)),$$

где $|\mathrm{SBS}^0 a_i|$ — мощность множества $\mathrm{SBS}^0 a_i$, $\exists$ — квантор существования, а $\mathrm{SBS}^0 a_i - a_j$ — разность двух множеств: $\mathrm{SBS}^0 a_i$ и множества, представленного единичным элементом a_j.

Легко видеть, что длина кода строки символов x определяется числом символов исходной строки, для которых выполняется правило 2), и плюс начальный символ.

Пользуясь данными правилами кодирования и ТВ-структурой (рис. 2), получим для слов, заданных конечным множеством X, следующие коды. Множеству слов {*кос'ить, кос'ишь, кос'ит, ко'сим, кос'ят, кос'а, кос'ы, кос'ой*} соответствует код, КО, множеству {*к'осишь, к'осит, к'оси, к'осят, к'осы, к'ос*} — код К, слову: *кош'у*: — код К, слову *п'ара* — код П'А, словам {*п'ора, п'оры*} — код П'О, слову *пор'а* — код ПО, слову *н'ос* — код Н, слову *к'озырь* — код К'З и слову *к'орь* — код К'Р.

Непосредственное использование кодов слов при переходе от поверхностной структуры предложения к глубинной невозможно по той причине, что коды слов будут изменяться при расширении ТВ-структуры. Покажем это на примере. Пусть задано конечное множество слов X = {*приносить* * *что-делать, просить* * *что-делать, приходить* * что-делать, пилить * что-делать, придавать * что-делать, носить * что-делать, давать * что-делать. Ударения в словах не указаны, поскольку они не влияют в данном случае на коды слов. Проследим, как изменяется код слова *приносить* по мере добавления в ТВ-структуру других слов из заданного множества X. Соответствующая этим словам ТВ-структура отражена на рис. 3, а результаты изменения кодов под воздействием первых пяти входных объектов даны ниже. Входные объекты: приносить, просить, приходить, пилить, придавать

Коды слова *приносить*: П ПИ ПИН ПРИН ПРИН

Как видно, код слова *приносить* меняется с каждым новым словом, не считая слова *придавать*. Для того чтобы коды, используемые в глубинной структуре, не изменялись, введем множество C, элементами которого будут коды слов, полученные с помощью ТВ-структуры. В глубинной структуре предложения кодами слов служат их номера в множестве C. Перед номером в качестве разделителя ставится символ *. Код слова *приносить* обозначается *1, *просить* — *2 и т. д. После ввода в ТВ-структуру слова *приходить*, множество C имеет вид: C = {ПИН, ПО, ПИХ}, после добавления слова *придавать* —

$C = \{$ПРИН, ПРО, ПРИХ, ПИ, ПРИД$\}$, коды этих слов в глубинной структуре записываются соответственно $*1$, $*2$, $*3$, $*4$, $*5$.

Итак, добавление новых слов в ТВ-структуру может привести к изменению кодов слов, ранее включенных в ТВ-структуру. Но при этом имеет место интересная закономерность, чем больше слов содержится в ТВ-структуре, тем меньше кодов требуют коррекции после ввода в ТВ-структуру нового слова.

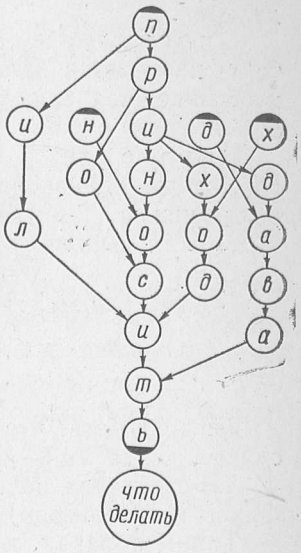

Рис. 3

Определение 8. Пусть код новой строки символов x, введенной в ТВ-структуру, будет $\mathrm{cod}(x)$. Тогда эта строка символов не приведет к коррекции кодов строк, уже записанных в ТВ-структуре, если имеет место следующее утверждение:

$$((a_i \leftarrow \mathrm{cod}(x)) \,\&\, (a_i \overset{0}{\underset{\mathrm{cod}(x)}{\rightarrow}} e)) \vdash (\exists a_j \leftarrow x \times$$

$$\times ((a_j \overset{0}{\underset{x}{\rightarrow}} a_i) \,\&\, (|\,\mathrm{SBS}^0 a_j\,| > 2)).$$

Включение в ТВ-структуру слова *придавать* не привело к коррекции кодов уже записанных слов по той причине, что $SBS^0_{\text{И}} = \{$Н, Х, Д$\}$, т. е. мощность этого множества больше двух.

Еще одну особенность кодирования слов продемонстрируем на следующем примере. Пусть задано множество $X = \{$*колбаса* $*$ $*$ *она* $*$ *что* $*$ *колба* $*$ *она* $*$ *что*$\}$. Легко представить соответствующую элементам множества X ТВ-структуру и коды этих слов: КА и КА, т. е. cod (*колбаса*) = КА и cod (*колба*) = КА. Буква А в коде слова *колба* отличается от буквы А в коде слова *колбаса* тем, что она является конечной. Отмечая эту букву как А', мы тем самым отличим код слова *колбаса* — КА' от кода первого слова — КА, т. е. в первом случае будем писать cod (*колбаса*) = КА, а во втором — cod (*колба*) = КА'. В общем случае для строки x это записывается следующим образом:

$$\left((a_i \leftarrow \mathrm{cod}(x)) \,\&\, (a_i \overset{0}{\underset{x}{\rightarrow}} e)\right) \vdash \mathrm{cod}'(x).$$

Рассмотренная система кодирования слов позволяет сохранять и фиксировать семантическую информацию, которая вносится в значение слова его приставкой. Характерно, что по мере расширения словаря системы ДЕСТА, а следовательно, и росте ТВ-структуры, происходит коррекция кодов, что в конце концов приведет к включению приставок в коды слов. Коды слов *при-*

носить, *приходить* и *придавать* включают в себя, в качестве начальных букв, приставку *при* (рис. 3). Умение системы ДЕСТА автоматически выделять приставку и связывать с ней семантическую информацию позволит приблизиться к решению задачи образования с помощью приставок новых слов. Для формального определения понятия приставка введем следующие определения.

Определение 9. Если x_1, x_2 и x_1x_2 — строки символов то x_2 называется *левым частным от деления* строки x_1x_2 на x и обозначается как $x_1 \setminus x_1x_2$, а x_1 называется *правым частным от деления* строки x_1x_2 на x_2 и обозначается через x_1x_2 / x_2.

Определение 10. Пусть x_1 и x_2 строки символов, сопровождаемые одинаковой морфологической информацией, которые включены в ТВ-структуру. Тогда *приставку* определим как x_1/x_2, если выполняется следующее условие:

$$\exists a_i \leftarrow x_1 \exists a_j \leftarrow x_2 \left((a_i = a_j) \,\&\, \left(a_j \overset{0}{\underset{x_2}{\rightarrow}} e \right) \,\&\, (\mathrm{SBS}^0 a_j = \mathrm{SBS}^0 a_i) \,\&\, \right.$$

$$\left. \&\, ((a_j \setminus x_2) \leftarrow x_1) \right).$$

Легко видеть, что $(x_1/x_2) \leftarrow \mathrm{cod}(x_1)$ будем иметь место в том случае, когда $\forall a_i \leftarrow (x_1/x_2) (|SBS^0 a_i)| > 1)$.

ТВ-структура позволяет также автоматически выделять суффиксы и использовать их при словообразовании.

Таким образом, кодирование слов в ТВ-структуре позволяет не только получать одинаковые коды различных словоформ, но и предоставляет возможность выделить приставки и суффиксы для использования их при словообразовании.

Список литературы: 1. *Ловицкий В. А.* Структурный подход к решению морфологических задач. — Проблемы бионики, 1980, вып. 25, с. 39—43. 2. *Глозман Ж. М.* Исследование нарушения, лингвистического отношения к слову при афазии. — Психологические исследования, 1974, вып. 6, с. 77—87.

Поступила 22 февраля 1979 г.