УДК 510.62

В. Я. ТЕРЗИЯН, канд. техн. наук, *И. И. ПОПКОВ*

ФОРМАЛИЗАЦИЯ ПРОЦЕССА УСТРАНЕНИЯ МНОГОЗНАЧНОСТИ СИНТАКСИЧЕСКОГО РАЗБОРА ЕСТЕСТВЕННОЯЗЫКОВОГО ВЫСКАЗЫВАНИЯ. Сообщение 1.

Анализ естественноязыковых текстов на этапе синтаксического разбора сталкивается с рядом трудностей, обусловленных прежде всего тем, что из-за синонимии различного рода и других причин возникает многозначность синтаксического представления естественноязыковых высказываний (ЕЯВ). Устранение многознач-

ности на этом этапе достигается применением синтаксического фильтра, работающего с учетом ограничений для входного ЕЯВ. Формализация процесса, характеризующего работу этого фильтра, позволяет эффективно анализировать простые ЕЯВ и удалять неподходящие варианты синтаксического разбора путем введения новых ограничивающих правил на более высоких уровнях анализа.

Рассмотрим входное простое ЕЯВ. Каждая словоформа рассматриваемого предложения $S_i \in S$, где S — множество всех возможных словоформ словаря. Обозначим через z мощность множества S. Для установления факта принадлежности каких-либо словоформ из множества S конкретному рассматриваемому высказыванию введем предикат $WORD(i)$. Данный предикат определен на множестве индексов всех словоформ и является предикатом принадлежности словоформы S_i словаря входному ЕЯВ:

$$WORD(i) \begin{cases} \text{1, } S \text{ принадлежит входному высказыванию} \\ \text{0, } S \text{ не принадлежит входному высказыванию,} \end{cases}$$

где $i=1, \ldots z$ (z — мощность словаря).

В результате морфологического анализа ЕЯВ каждой словоформе соответствует один или несколько наборов признаков, позволяющих, в конечном итоге, устанавливать синтаксические связи в предложении. Каждый набор признаков представляет собой вектор, состоящий из n элементов $p^1, \ldots, p^n$, являющихся элементами подмножеств $P_1, P_2, \ldots, P_n$ множества M значений всех признаков словоформ. В работе [1] в качестве набора признаков выделяется местоимение (определяет род, число), вопрос, на который отвечает словоформа, и предлог, имеющий отношение к данной словоформе. Например, в высказывании «Белокурый мальчик едет на велосипеде» вектор признаков для каждой словоформы этого ЕЯВ будет записан в виде p^1, p^2, p^3, где p^1 — элемент подмножества всех вопросительных слов: $P_1=\{$кто, что, ...$\}$, p^2 — элемент подмножества местоимений: $P_2=\{$он, она, оно, они$\}$, p^3 — элемент подмножества всех предлогов: $P_3=\{$в, на, к, по, ...$\}$. Каждое подмножество P_i содержит нулевой элемент 0, который входит в вектор признаков в том случае, если соответствующий элемент отсутствует для конкретной словоформы S_i.

Таким образом, для нашего примера вектор признаков для словоформы «белокурый» будет иметь вид: (какой, он, 0); для словоформы «Мальчик»: (кто, он, 0); «едет»; (что—делать, он, 0); «велосипеде»: (чем, он, на). Некоторые словоформы могут иметь несколько векторов признаков: «рабочий»: (какой, он, 0) и (кто, он, 0), что в конечном счете приводит к нескольким вариантам синтаксического разбора высказывания.

Рассмотрим множество векторов признаков как декартово произведение подмножеств одного класса: $P_1, P_2, \ldots, P_n$: $V=P_1 \times \times P_2 \times \ldots \times P_n$, где V — множество векторов признаков.

Пусть множество $P_1=\{p_1^1, p_2^1, \ldots, p_{k_1}^1\}$, где k_1 — количество элементов данного подмножества. Соответственно: $P_2=\{p_1^2,$

$p_2^2, \ldots, p_{k_2}^2\}$ и т. д.: $P_n = \{p_1^n, p_2^n, \ldots p_n^n\}$. Тогда мощность множества $V{:}K = k_1 k_2 \ldots k_n$.

Пронумеруем элементы множества векторов признаков следующим образом: $v_1 = (p_1^1, p_1^2, \ldots, p_1^n)$; $v_2 = (p_1^1, p_1^2, \ldots, p_2^n) \ldots v_{k_1} = (p_1^1, p_1^2, \ldots, p_{k_1}^n)$; $v_{k_1+1} = (p_1^1, p_1^2, \ldots p_2^{n-1}, p_1^n) \ldots v_{k_n} = (p_{k_1}^1, p_{k_2}^2, \ldots, p_{k_n}^n)$.

На этапе морфологического анализа входного высказывания возникает задача поставить в соответствие каждой словоформе один или несколько векторов признаков из множества V. Другими словами, необходимо определить истинность предиката $MORF\,(i, j)$, который задается так:

$$MORF(i, j) = \begin{cases} 1, \text{ если } v_j \text{ соответствует } S_i; \\ 0, \text{ если } v_j \text{ не соответствует } S_i, \end{cases}$$

где $i = 1, \ldots, z$; $j = 1, \ldots, K$.

Синтаксические связи между словоформами входного высказывания, определенные на уровне соответствующих векторов признаков, являются направленными и необходимы для построения дерева синтаксического разбора — формируются с помощью предиката $LINK$, задаваемого следующим образом:

$$LINK(i, j) = \begin{cases} 1, \text{ если } v_i \text{ имеет синтаксическую связь,} \\ \text{направленную от } v_i \text{ к } v_j; \\ 0, \text{ в противном случае.} \end{cases}$$

Истинность предиката $LINK(i, j)$ определяется с помощью валентностей [2], которые формируются автоматически на этапе обучения системы, процедурных правил согласования [1] и др.

Целью синтаксического анализа является получение синтаксического дерева анализируемого ЕЯВ. Поэтому необходимо привести условия, которым должно удовлетворять дерево синтаксического разбора анализируемой структуры. Введем следующие определения: *корневой* называется словоформа, не имеющая ни одной подходящей к ней синтаксической связи; *некорневой* называется словоформа, имеющая одну подходящую синтаксическую связь. В правильно построенном синтаксическом дереве должна быть одна корневая словоформа, а остальные — некорневые. Кроме того, каждая словоформа должна иметь в своем составе только одну группу признаков (вектор).

При обработке реальных ЕЯВ возникают многозначности на этапе построения синтаксической структуры, удовлетворяющей указанным требованиям. Целью анализа на данном этапе является построение всех вариантов правильных синтаксических деревьев. Например, при анализе высказывания «Красное мороженое стекло в чашку», словоформа «стекло» может рассматриваться как су-

ществительное (причем и в именительном, и в винительном падежах), а также как глагол (прошедшее время среднего рода от «стекать»), «мороженое» может интерпретироваться как существительное (продукт питания) в именительном или винительном падежах, либо как прилагательное (от «мороженый). Следовательно, отдельные словоформы могут иметь более одного варианта векторов признаков, что после применения предиката *LINK(i, j)* приводит к многозначности синтаксической структуры.

При формировании синтаксической структуры входного ЕЯВ подразумевается истинность предикатов *WORD(i)*, *MORF(i, j)* и *LINK(i, j)*, позволяющих соотнести с анализируемым высказыванием соответственно: множество всех словоформ *S*, множество всех векторов признаков *V* с конкретными словоформами и также множество состоящих в синтаксическом отношении конкретных векторов признаков, позволяющих создать исходную структуру рассматриваемого ЕЯВ. Следовательно, многозначность анализируемой синтаксической структуры возникает из-за того, что некоторые словоформы ЕЯВ имеют более одного вектора признаков v_j, а также из-за многозначности, возникающей вследствие применения предиката *LINK(i, j)*.

Синтаксический разбор входного высказывания можно представить в виде дизъюнкции всех вариантов синтаксических деревьев:

$$R_i = WORD(i)\,\tilde{F}_i \bigwedge_{r=1}^{z} F_r, \qquad (1)$$

где $\tilde{F}_i$ — предикат, характеризующий корневую словоформу, который задается следующим образом:

$$\tilde{F}_i = \begin{cases} 1, & \text{если } i\text{-я словоформа является корневой;} \\ 0, & \text{если } S_i \text{ не является корневой.} \end{cases}$$

Соответственно, F_r — предикат, характеризующий некорневую словоформу

$$F_r = \begin{cases} 1, & \text{если } S_i \text{ является некорневой;} \\ 0, & \text{если } S_i \text{ не является некорневой.} \end{cases}$$

R_i — выражение для i-го варианта синтаксического дерева.

Отметим, что словоформа может быть: корневой, некорневой, а также неправильной — иметь более одного вектора признаков (т. е. в общем случае $F_r \neq \tilde{F}_i$, и наоборот; $r=1, ..., z$; $i=1, ..., z$).

Но выражение (1) имеет в своем составе предикат *WORD(i)*, который учитывает принадлежность i-й словоформы анализируемому ЕЯВ, но не проверяет выполнимость этого же условия для F_r. В конъюнкции необходимо, чтобы все словоформы F_r, не входящие в данное ЕЯВ, не учитывались, т. е. соответственно равнялись 1. Это выполняется в следующем случае: $F_r \vee \overline{WORD(r)}$. Когда F_r не будет входить в анализируемое ЕЯВ, *WORD(r)* будет

равен 0, соответственно $\overline{WORD(r)}=1$, которая поглощает F_r. Также надо учесть тот факт, что индексы r и i не могут быть равны (т. е. одна словоформа в рассматриваемом варианте дерева не может быть одновременно корневой и некорневой). Поэтому для учета этого условия надо ввести предикат равенства $D(r, i)$:

$$D(r, i)=\begin{cases}1, & \text{если } r=i,\\ 0, & \text{если } r\neq i\end{cases}$$

и использовать поглощение F_r в дизъюнкции единицей. Таким образом, для некорневой словоформы имеем выражение $F_r\vee \vee\overline{WORD(r)}\vee D(r, i)$, которое учитывает все ранее приведенные условия. Окончательная структура входного высказывания может быть записана уравнением

$$\bigvee_{i=1}^{V}(WORD(i)\,\tilde{F}_i\bigwedge_{r=1}^{z}(F_r\vee\overline{WORD}(r)\vee D(r, i))=1, \qquad (2)$$

где z — общее количество словоформ множества S.

Целью решения этого логического уравнения является определение связей синтаксического дерева для данного ЕЯВ. Отсюда следует, что необходимо выразить предикаты $\tilde{F}_i$ и F_r через переменные, характеризующие связи между векторами признаков дерева синтаксического разбора. Поэтому введем предикат $SYNT(i, j, k, l)$, определенный на множестве индексов i, j, k, l, удовлетворяющих следующим условиям:

$$WORD(i)=1;\quad WORD(k)=1;\quad MORF(i, j)=1;$$
$$MORF(k, l)=1;\quad LINK(j, l)=1;\quad D(i, k)=0.$$

Определим предикат $SYNT(i, j, k, l)$, характеризующий направленную синтаксическую связь между векторами признаков различных словоформ анализируемого высказывания:

$$SYNT(i, j, k, l)=\begin{cases}1, \text{ если синтаксическая связь, направленная от } v_j, \text{ соответствующего словоформе } S_i, \text{ к вектору } v_l, \text{ соответствующего словоформе } S_k, \text{ включается в синтаксическое дерево;}\\ 0. \text{ если данная связь не включается в синтаксическое дерево.}\end{cases}$$

С учетом введенного предиката запишем выражения для $\tilde{F}_i$ и F_r. При составлении окончательных уравнений для корневой и некорневой словоформ необходимо использовать понятия «исключения всех подходящих» либо «отходящих» связей и др., целесообразно ввести выражения для записи дизъюнкции и конъюнкции для подходящих и отходящих связей к v_j, который соответствует словоформе S_i. Дизъюнкцию всех отходящих от v_j связей (слово-

формы S_i) $OUTD(i, j)$ необходимо записать путем просмотра всех словоформ анализируемого ЕЯВ (за исключением S_i, что легко учитывается введением предиката равенства, аналогично (2)) на наличие синтаксических связей с v_j. Так как при составлении $OUTD(i, j)$ рассматривается все множество S и все множество V, то, естественно, необходимо учитывать истинность предикатов $WORD(k)$ и $MORF(k, m)$:

$$OUTD(i, j) = \bigvee_{k=1}^{z} (WORD(k)\,\overline{D}(i, k) \cdot \bigvee_{m=1}^{K} (MORF(k, m) \wedge$$

$$\wedge LINK(j, m)\,SYNT(i, j, k, m))), \qquad (3)$$

где z — мощность множества S; K — мощность множества V.

Предикат $\overline{D}(i, k)$ позволяет исключить из рассмотрения S_i, так как при $i=k$, $\overline{D}(i, k)=0$ и в общей дизъюнкции по k соответствующий дизъюнкт будет равен также 0.

Соответственно для записи конъюнкции всех синтаксических связей, отходящих от v_j словоформы S_i, выражение $OUTK(i, j)$ имеет вид

$$OUTK(i, j) = \bigwedge_{k=1}^{z} (\overline{WORD}(k) \vee D(i, k) \vee$$

$$\vee \bigwedge_{m=1}^{K} (\overline{MORF}(k, m) \vee \overline{LINK}(j, m) \vee SYNT(i, j, k, m))). \qquad (4)$$

Выражения для подходящих к v_j связей записываются аналогично (3) и (4). Для дизъюнкции подходящих синтаксических связей

$$IND(i, j) = \bigvee_{k=1}^{z} (WORD(k) \cdot \overline{D}(i, k) \cdot \bigvee_{m=1}^{K} (MORF(k, m) \wedge$$

$$\wedge LINK(m, j)\,SYNT(k, m, i, j))). \qquad (5)$$

Конъюнкция подходящих синтаксических связей запишется в виде

$$INK(i, j) = \bigwedge_{k=1}^{z} (\overline{WORD}(k) \vee D(i, k) \vee$$

$$\vee \bigwedge_{m=1}^{K} (\overline{MORF}(k, m) \vee \overline{LINK}(m, j) \vee SYNT(k, m, i, j))). \qquad (6)$$

В дальнейшем для записи логических выражений для корневой и некорневой словоформ, помимо выражений (3) — (5), необходимо использовать аналогичные выражения, с той лишь разницей, что предикат $SYNT$, характеризующий связи в синтаксических деревьях, должен входить с отрицанием: $\overline{SYNT}(i, j, k, l)$. Такие выражения будем обозначать соответственно: $OUTD^*(i, j)$ **(7)**, $OUTK^*(i, j)$ **(8)**, $IND^*(i, j)$ **(9)**, $INK^*(i, j)$ **(10)**. С учетом этого

запишем выражения для F_i и F_r относительно предиката $SYNT(i, j, k, l)$.

При составлении уравнения для корневой словоформы $\tilde{F}_i$ возникают следующие условия: для каждого вектора признаков v_j, соответствующих $\tilde{F}_i$, все связи записываем без отрицания через дизъюнкцию; при этом все подходящие к данной словоформе связи записываем через логическое отрицание (т. е. $\overline{SYNT}(k, l, i, j)$, $k \neq i$) и все отходящие от словоформы $\tilde{F}_i$ связи, кроме соответствующих вектору признаков v_j, также записываются через отрицание (т. е. $\overline{SYNT}(i, j, k, l)$, $i \neq k$, $j \neq l$. Таким образом, для корневой словоформы $\tilde{F}_i$ имеем следующее выражение:

$$\tilde{F}_i = \bigvee_{j=1}^{K} (MORF(i, j)\, OUTD(i, j) \vee \bigwedge_{t=1}^{K} (\overline{MORF}(i, t) \vee \vee INK^*(i, t) \cdot (OUTK(i, t) \vee D(t, j))), \quad (11)$$

где K — мощность множества векторов признаков.

Так как при составлении выражения для корневой словоформы анализируются все элементы множества V, то предикат $MORF(i, j)$ позволяет исключить из дизъюнкции по j те вектора признаков, которые не относятся к S_i анализируемого высказывания; $\overline{MORF}(i, j)$ проводит аналогичную операцию для конъюнкции синтаксических связей.

При записи выражения, характеризующего некорневую словоформу, используем следующее правило: все связи, подходящие к конкретному вектору v_j, соответствующему словоформе S_i, должны записываться через дизъюнкцию конъюнкций, в которых поочередно каждая подходящая к v_j связь записывается без отрицания, а остальные подходящие связи записываются с отрицанием, что соответствует предикатам $SYNT(k, l, i, j)$ и $\overline{SYNT}(k, l, i, j)$. Данное правило можно представить выражением

$$ENOT(i, j) = \bigvee_{k=1}^{z} (WORD(k) \cdot \overline{D}(k, i) \cdot \bigvee_{t=1}^{K} (MORF(k, t) \wedge \wedge LINK(t, j)\, SYNT(k, t, i, j) \cdot \bigwedge_{p=1}^{K} (MORF(k, \Lambda) \wedge \wedge LINK(p, j) \cdot (\overline{SYNT}(k, p, i, j) \vee D(p, t))))), \quad (12)$$

где z — мощность множества всех словоформ S;
K — мощность множества векторов признаков.

Так как при составлении $ENOT(i\, j)$, рассматриваются все элементы множеств S и V, предикаты $WORD(k)$ и $MORF(k, t)$ позволяют исключить из рассмотрения все словоформы и вектора признаков, не относящиеся к данному ЕЯВ. Предикат равенства $D(p, t)$ позволяет исключить из (12) все синтаксические отношения, возникающие внутри словоформы S_i.

С учетом (12) запишем полностью выражение для некорневой словоформы, принимая во внимание, что кроме связей, которые описывает $ENOT(i, j)$, через конъюнкцию записываются все подходящие и отходящие к другим $v_t (t \neq j)$ соответствующие словоформе S_i синтаксические связи. Дизъюнкция по всем векторам признаков данной некорневой словоформы, с учетом приведенных условий, дает требуемое выражение для S_i анализируемого ЕЯВ:

$$F_r = \bigvee_{j=1}^{K} (MORF(r, j)\, ENOT(r, j) \cdot \bigwedge_{t=1}^{K} (\overline{MORF}(r, t) \vee$$

$$\vee INK^*(r, t)\, OUTK^*(r, t) \vee D(t, j)))). \qquad (13)$$

Конъюнкция по t фиксирует все синтаксические связи, подходящие и отходящие от словоформы S_r с отрицанием (т. е. $\overline{SYNT(i, j, k, l)}$ и $SYNT(k, l, i, j)$, а предикат $D(t, j)$ позволяет исключить из рассмотрения связи, соответствующие текущему вектору признаков v_j, входящему в S_i.

Синтаксический анализ реальных ЕЯВ во многих случаях позволяет получить единственное дерево синтаксического разбора. В случаях, когда возникает несколько вариантов деревьев, дальнейшая обработка полученных структур может проводиться с учетом дополнительных ограничений. Требование выполнения условий проективности деревьев, «правила соседства» и т. д. означает удаление неподходящих вариантов синтаксических деревьев.

Список литературы: 1. *Ловицкий В. А.* Диалоговая естественноязыковая система принятия решений. Х., 1981. 110 С. 2. *Терзиян В. Я.* Принципы организации анализа естественноязыкового высказывания в системах общения пользователей с ЭВМ. Сообщ. 2//Пробл. бионики. 1985. Вып. 35. С. 17—24.

Поступила в редколлегию 14.09.90