УДК 510.62

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

ФОРМАЛИЗАЦИЯ ПРОЦЕССА УСТРАНЕНИЯ МНОГОЗНАЧНОСТИ СИНТАКСИЧЕСКОГО РАЗБОРА ЕСТЕСТВЕННОЯЗЫКОВОГО ВЫСКАЗЫВАНИЯ. Сообщение 2

Приведенный в [1] метод построения синтаксических деревьев, позволяющий получать несколько вариантов синтаксических структур, удовлетворяющих заданным условиям, эффективно работает при анализе простых распространенных предложений. Во многих случаях позволяет получить единственную древовидную синтаксическую структуру. Рассмотрим конкретный пример. Пусть дано естественноязыковое высказывание (ЕЯВ): «Красное мороженое стекло в чашку». Проведем синтаксический анализ этого высказывания и построим структуру данного ЕЯВ, используя результаты работы [1].

Предположим, что словарь (множество S) помимо других, также включает в себя следующие элементы: $S=\{...,$ красное, ..., мороженое, ..., стекло, ..., чашку, ...$\}$. Истинность предиката $WORD(i)$ для рассматриваемого ЕЯВ позволяет включить в анализ только те элементы множества S, которые содержатся в данном предложении. Пусть словоформы входного ЕЯВ имеют следующие номера в S: i_1, i_2, i_3, i_4. Таким образом, $WORD(i_1)=1$; $WORD(i_2)=1$; $WORD(i_3)=1$; $WORD(i_4)=1$. В рассматриваемом случае данный предикат для индексов всех остальных элементов равен нулю.

Введем множество векторов признаков V, состоящее из элементов: $V=\{...,$ (какой, оно, 0), ..., (что — в, оно, 0), ..., (что, оно, 0), ..., (что — сделало, 0, 0), ..., (куда, 0, в), ...$\}$. Вектора признаков в данном случае формируются из множеств: $P_1=\{0, ...,$ какой, ..., что, ..., что — в, ..., что — сделало, ..., куда, ...$\}$; $P_2=\{0,$ он, она, оно, они$\}$; $P_3=\{0, ...,$ в, ...$\}$.

Присвоим приведенным элементам множества V соответственно индексы: j_1, j_2, j_3, j_4, j_5. Следовательно, анализируя истинность предиката $MORF(i, j)$, можно установить, какие векторы признаков характеризуют каждую S_i входного предложения:

$$MORF(i_1, j_1)=1;\ MORF(i_2, j_1)=1;\ MORF(i_2, j_2)=1;\ MORF(i_2, j_3)=1;\ MORF(i_2, j_4)=1;\ MORF(i_3, j_3)=1;\ MORF(i_3, j_4)=1;$$

$$MORF(i_4, j_5)=1.$$

Значения этого предиката по всем остальным индексам в рамках входного ЕЯВ равны нулю.

Предикат $LINK(i, j)$ позволяет установить синтаксическую взаимосвязь векторов признаков множества V:

$$LINK(j_2, j_1)=1;\ LINK(j_3, j_1)=1;$$

$$LINK(j_4, j_3)=1;\ LINK(j_4, j_5)=1.$$

Для всех остальных значений индексов векторов признаков, характеризующих анализируемое ЕЯВ, предикат $LINK(i, j)=0$.

Найдем $OUTD(i, j)$, учитывая значение индексов i и j, при которых предикаты $WORD(i)$, $MORF(i, j)$, $LINK(i, j)$ принимают истинное значение: $OUTD(i_1, j_1)=0$; $OUTD(i_2, j_1)=0$; $OUTD(i_2, j_2)=SYNT(i_2, j_2, i_1, j_1)$; $OUTD(i_2, j_3)=SYNT(i_2, j_3, i_1, j_3)$; $OUTD(i_3, j_3)=SYNT(i_3, j_3, i_1, j_1) \vee SYNT(i_3, j_3, i_2, j_1)$; $OUTD(i_3, j_4)=SYNT(i_3, j_4, i_2, j_3) \vee SYNT(i_3, j_4, i_4, j_5)$; $OUTD(i_4, j_5)=0$.

Найдем, чему равна конъюнкция всех отходящих синтаксических связей от каждой словоформы рассматриваемого ЕЯВ: для 1-й словоформы: $OUTK(i_1, j_1)=0$; для 2-й словоформы:

$$OUTK(i_2, j_1)=0;\ OUTK(i_2, j_2)=SYNT(i_2, j_2, i_1, j_1);$$

$OUTK(i_2, j_3) = SYNT(i_2, j_3, i_1, j_1)$; для 3-й словоформы:

$OUTK(i_3, j_3) = SYNT(i_3, j_3, i_2, j_1) \wedge SYNT(i_3, j_3, i_1, j_1)$;

$OUTK(i_3, j_4) = SYNT(i_3, j_4, i_2, j_3) \wedge SYNT(i_3, j_4, i_4, j_5)$;

для 4-й словоформы: $OUTK(i_4, j_5) = 0$.

Определим INK^* для словоформ данного ЕЯВ: для 1-й словоформы $INK^*(i_1, j_1) = \overline{SYNT}(i_2, j_2, i_1, j_1) \wedge \overline{SYNT}(i_2, j_3, i_1, j_1) \wedge \overline{SYNT}(i_3, j_3, i_1, j_1)$; для 2-й словоформы: $INK^*(i_2, j_1) = \overline{SYNT}(i_3, j_3, i_2, j_1)$; $INK^*(i_2, j_2) = 1$; $INK^*(i_2, j_3) = \overline{SYNT}(i_3, j_4, i_2, j_3)$; для 3-й словоформы: $INK^*(i_3, j_3) = 1$; $INK^*(i_3, j_4) = 1$; для 4-й словоформы: $INK^*(i_4, j_5) = \overline{SYNT}(i_3, j_4, i_4, j_5)$.

Определим для каждой словоформы $\tilde{F}_l$: $F_{t_1} = 0$; $F_{t_2} = SYNT(i_2, j_2, i_1, j_1)\,\overline{SYNT}(i_3; j_3, i_2, j_1)\,\overline{SYNT}(i_3, j_4, i_2, j_3)\,\overline{SYNT}(i_2, j_3, i_1, j_1) \vee SYNT(i_3, j_3, i_2, j_1) \wedge \overline{SYNT}(i_2, j_2, i_1, j_1)\,\overline{SYNT}(i_3, j_4, i_2, j_3) \cdot \tilde{F}_{t_3} = (SYNT(i_3, j_3, i_2, j_1) \vee SYNT(i_3, j_3, i_1, j_1)) \wedge \overline{SYNT}(i_3, j_4, i_2, j_3)\,\overline{SYNT}(i_3, j_4, i_4, j_5) \vee (SYNT(i_3, j_4, i_2, j_3) \vee \wedge SYNT(i_3, j_4, i_4, j_5)) \vee \overline{SYNT}(i_3, j_3, i_2, j_1)\,\overline{SYNT}(i_3, j_3, i_1, j_1) \cdot \tilde{F}_{t_4} = 0$.

Для того, чтобы найти выражение для некорневой словоформы, необходимо найти значение выражения $ENOT(i, j)$ для всех словоформ данного ЕЯВ: для 1-й словоформы: $ENOT(i_1, j_1) = SYNT(i_2, j_2, i_1, j_1)\,\overline{SYNT}(i_2, j_3, i_1, j_1)\,\overline{SYNT}(i_3, j_3; i_1, j_1) \vee \overline{SYNT}(i_2, j_3; i_1, j_1)\,\overline{SYNT}(i_2, j_2, i_1, j_1) \wedge SYNT(i_3, j_3, i_1, j_1) \vee SYNT(i_3, j_3, i_1, j_1)\,\overline{SYNT}(i_2, j_2, i_1, j_1)\,\overline{SYNT}(i_2, j_3, i_1, j_1)$; для 2-й словоформы: $ENOT(i_2, j_1) = SYNT(i_3, j_3, i_2, j_1)$; $ENOT(i_2, j_2) = 0$; $ENOT(i_2, j_3) = SYNT(i_3, j_4, i_2, j_3)$; для 3-й словоформы данные выражения равны нулю; для 4-й словоформы $ENOT(i_4, j_5) = SYNT(i_3, j_4, i_4, j_5)$.

Запишем выражения для некорневых словоформ F_l: $F_{t_1} = SYNT(i_2, j_2, i_1, j_1)\,\overline{SYNT}(i_2, j_3, i_1, j_1) \wedge \overline{SYNT}(i_3, j_3, i_1, j_1) \vee \vee SYNT(i_2, j_3, i_1, j_1) \wedge \overline{SYNT}(i_2, j_2, i_1, j_1)\,\overline{SYNT}(i_3, j_3, i_1, j_1) \vee \vee SYNT(i_3, j_3, i_1, j_1)\,\overline{SYNT}(i_2, j_2, i_1, j_1) \wedge \overline{SYNT}(i_2, j_3, i_1, j_1)$; $F_{t_2} = SYNT(i_3, j_3, i_2, j_1) \wedge \overline{SYNT}(i_3, j_4, i_2, j_3)\,\overline{SYNT}(i_2, j_2, i_1, j_1) \wedge \overline{SYNT}(i_3, j_3, i_2, j_1)$; $F_{t_3} = 0$; $F_{t_4} = SYNT(i_3, j_4, i_4, j_5)$.

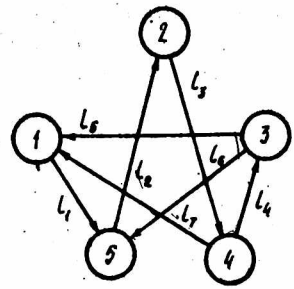
Таким образом, уравнение синтаксической структуры данного ЕЯВ будет иметь вид $F_{t_3}F_{t_1}F_{t_2}F_{t_4} = 1$. $((SYNT(i_3, j_3, i_2, j_1) \vee SYNT(i_3, j_3, i_1, j_1))\,\overline{SYNT}(i_3, j_4, i_2, j_3) \wedge \overline{SYNT}(i_3, j_4, i_4, j_5) \vee \overline{SYNT}(i_3, j_4, i_2, j_3) \vee SYNT(i_3, j_4, i_4, j_5)\,\overline{SYNT}(i_3, j_3, i_2, j_1) \wedge \overline{SYNT}(i_3, j_3, i_1, j_1))\,(SYNT(i_2, j_2, i_1, j_1)\,\overline{SYNT}(i_2, j_3, i_1, j_1) \times$

$\times \overline{SYNT}(i_3, j_3, i_1, j_1) \vee SYNT(i_2, j_3, i_1, j_1) \wedge \overline{SYNT}(i_2, j_2, i_1, j_1) \times$
$\times \overline{SYNT}(i_3, j_3, i_3, j_1) \vee SYNT(i_3, j_3, i_1, j_1) \overline{SYNT}(i_2, l_2, i_1, j_1) \times$
$\times \overline{SYNT}(i_2, j_3, i_1, j_1) (SYNT(i_3, j_3, i_2, j_1) \overline{SYNT}(i_3, j_4, i_2, j_3) \times$
$\times \overline{SYNT}(i_2, j_2, i_1, j_1) \overline{SYNT}(i_2, j_3, i_1, j_1) \vee SYNT(i_3, j_4, i_2, j_3) \wedge$
$\wedge \overline{SYNT}(i_2, j_2, i_1, j_1) \overline{SYNT}(i_3, j_3, i_2, j_1) \wedge \overline{SYNT}(i_3, j_4, i_4, j_5) =$
$= 1.$

В итоге получаем $SYNT(i_2, j_2, i_1, j_1)$ $SYNT(i_2, j_3, i_1, j_1) \times$
$\times \overline{SYNT}(i_3, j_3, i_1, j_1) \overline{SYNT}(i_3, j_3, i_2, i_1) \wedge SYNT(i_3, j_4, i_2, j_3) \times$
$\times SYNT(i_3, j_4, i_4, j_5) = 1$. Таким образом, имеем $SYNT(i_2, j_2, i_1, j_1) = 0$; $SYNT(i_2, j_3, i_1, j_1) = 1$; $SYNT(i_3, j_3, i_1, j_1) = 0$; $SYNT(i_3, j_3, i_2, j_1) = 0$; $SYNT(i_3, j_4, i_2, j_3) = 1$; $SYNT(i_3, j_4, i_4, j_5) = 1$.

Необходимо отметить, что полученная структура проанализированного высказывания является однозначной. Истинность предикатов $SYNT(i, j, k, l)$ определяет направленную синтаксическую связь между словоформами выходной структуры.

Рассмотрим следующий пример абстрактной структуры (рисунок). Проведем анализ данной структуры и найдем все деревья, удовлетворяющие приведенным выше условиям. Для упрощения анализа применим метод частичного устранения многозначности синтаксического представления ЕЯВ, считая, что все предикаты принадлежности словоформы данному ЕЯВ равны единице, как и предикаты соответствия векторов признаков данной словоформе и др. (также, для простоты анализа считаем, что каждая словоформа высказывания имеет единственный вектор признаков). В результате анализа получается несколько древовидных синтаксических структур, полностью удовлетворяющих приведенным выше условиям: корень — словоформа 4:



$l_1 l_2 l_3 \overline{l_4} l_5 \overline{l_6} l_7 \vee \overline{l_1} l_2 \overline{l_3} l_4 \overline{l_5} l_6 l_7 \vee l_1 l_2 \overline{l_3} l_4 \overline{l_5} \overline{l_6} l_7$; корень — словоформа 5: $\overline{l_1} l_2 l_3 l_4 l_5 \overline{l_6} \overline{l_7} \vee \overline{l_1} l_2 l_3 l_4 \overline{l_5} l_6 l_7$; корень — словоформа 3: $l_1 l_2 l_3 \overline{l_4} l_5 \overline{l_6} l_7 \vee$
$\vee \overline{l_1} l_2 l_3 \overline{l_4} l_5 l_6 \overline{l_7} \vee \overline{l_1} l_2 l_3 \overline{l_4} \overline{l_5} l_6 l_7$; корень — словоформа 2: $\overline{l_1} l_2 l_3 l_4 \overline{l_5} l_6 l_7 \vee$
$l_1 \overline{l_2} l_3 l_4 l_5 \overline{l_6} \overline{l_7} \vee l_1 \overline{l_2} l_3 l_4 \overline{l_5} l_6 l_7 \vee \overline{l_1} l_2 l_3 l_4 l_5 l_6 l_7$; корень — словоформа 1: $l_1 l_2 l_3 \overline{l_4} \overline{l_5} \overline{l_6} l_7$.

Таким образом, после этапа предварительного анализа имеем 13 вариантов правильных синтаксических деревьев. В дальнейшем, следуя традиционным способам анализа, переходим на более высокий, семантический уровень для устранения многозначности. Потребуем от синтаксического дерева, чтобы выполнялось условие: ни одна синтаксическая связь не должна пересекаться

с другой в анализируемом ЕЯВ. В этом случае из 13 вариантов синтаксических деревьев остается лишь один: $l_1 l_2 l_3 l_4 l_5 l_6 l_7$. Причем данный вариант можно получить еще на этапе построения синтаксической структуры ЕЯВ, потребовав лишь выполнения условия непересечения синтаксических связей. Это условие является настолько мощным, что в рассматриваемом примере позволяет полностью устранить многозначность, исключая 12 вариантов синтаксических деревьев. Данное условие позволяет находить так называемые проективные синтаксические структуры.

Таким образом, проективным назовем дерево, в котором: 1) отсутствуют пересечения синтаксических связей; 2) ни одна синтаксическая связь не накрывает другую.

Если для синтаксического дерева выполняется лишь первое условие, а второе нет, то такое дерево назовем слабопроективным.

Для того чтобы записать условие непересечения связей, рассмотрим попарно все связи в анализируемом ЕЯВ (рисунок). Связь l_2 пересекается со связью l_7 и со связью l_5. Для того чтобы структура была слабопроективной, необходимо потребовать следующее: или отсутствуют l_2 или l_7, или l_2 и l_7, и отсутствуют связи l_2 или l_5. Аналогичное условие и для связей: l_3, l_6, l_7: необходимо отсутствие l_3 или l_6 и l_3 или l_7. Таким образом, условие для слабопроективной структуры запишется в виде

$$\bigwedge_{j=1}^{s} \bigwedge_{l=1}^{s} (\bar{l}_l \vee \bar{l}_j \vee C(l_l, l_j)) = 1, \qquad (1)$$

где l_i, l_j — синтаксические связи анализируемого ЕЯВ; C — количество связей в ЕЯВ. Предикат $C(l_i, l_j)$ — определяет пересечение синтаксических связей в анализируемом ЕЯВ:

$$C(l_l, l_j) = \begin{cases} 1, & \text{если } l_l \text{ и } l_j \text{ пересекаются,} \\ 0, & \text{если } l_l \text{ и } l_j \text{ не пересекаются.} \end{cases}$$

Для рассматриваемой структуры (рис. 1) условие слабопроективности выглядит слдующим образом:

$$(\bar{l}_2 \vee \bar{l}_5)(\bar{l}_2 \vee \bar{l}_7)(\bar{l}_3 \vee \bar{l}_5)(l_3 \vee \bar{l}_6) = 1. \qquad (2)$$

Умножая полученные 13 вариантов на это условие, в итоге находим единственный вариант синтаксического дерева, который не имеет в своей структуре пересекающихся синтаксических связей 5: $l_1 l_2 l_3 l_4 \bar{l}_5 l_6 \bar{l}_7$.

Требование от синтаксических деревьев выполнения условия проективности приводит к сильному сужению возможных вариантов синтаксических структур реальных высказываний. В связи с этим можно привести следующий аргумент в пользу такого ограничения, учитывая стоящие практические задачи анализа естественного языка: «В классе всевозможных деревьев подчинения можно выделить подкласс, который содержит подавляющее большинство «естественных» деревьев для предложений реальных языков

и, по-видимому, практически все «естественные» деревья выражений наиболее важных формализованных искусственных языков. Это класс так называемых проективных деревьев» [2].

На предварительном этапе частичного устранения многозначности синтаксического представления эффективными средствами минимизации являются требование построения правильной древовидной структуры и/или анализ синтаксической структуры на проективность. Причем для языков европейского типа возможен различный подход на предварительном этапе анализа: 1) составляются варианты синтаксических структур с учетом условий единственности корневой словоформы и т. д., а затем полученные варианты анализируются с точки зрения удовлетворения условию проективности; 2) вначале составляются варианты проективных синтаксических структур, а затем каждый из них рассматривается с позиции удовлетворения условиям правильной древовидной синтаксической структуре. Первый вариант предварительного анализа более эффективен для высказываний восточноевропейских языков (русский, польский, болгарский, украинский и т. д.), а второй вариант целесообразнее применять для западноевропейских языков (английский, немецкий, французский и т. д.) с более жестким порядком следования словоформ в предложении.

При анализе высказываний возможен и третий, совместный вариант решения задачи устранения многозначности синтаксического представления предложения: синтаксическое дерево высказывания составляется сразу с учетом и требований проективности структуры.

Предлагаемый способ минимизации возможен для реализации на ЭВМ в том случае, если он будет описан в терминах продукционного подхода, либо с помощью предикатов, определяющих принадлежности словоформ словарного множества анализируемому высказыванию, соответствие вектора признаков какой-либо словоформе данного ЕЯВ и т. д. Также необходимо определить предикат, устанавливающий наличие или отсутствие пересечения между парой синтаксических связей ЕЯВ, который определен на множестве индексов всех словоформ.

Содержательный смысл условий проективности и слабой проективности состоит в том, что слова, близкие синтаксически, близки и по положению в тексте. В научной и деловой прозе подавляющее большинство синтаксических деревьев являются проективными или слабопроективными. За некоторыми исключениями, непроективность в «деловом» тексте, скорее всего, характеризует слабую грамотность его сочинителя. Поэтому применение условий проективности синтаксического представления является важным требованием при анализе деловой прозы и позволяет эффективно устранять многозначность древовидной структуры анализируемого ЕЯВ.

Список литераутры: 1. *Терзиян В. Я., Попков И. И.* Формализация процесса устранения многозначности синтаксического разбора естественноязыкового высказывания. Сообщение 1//См. статью в настоящем сб. 2. *Гладкий А. В.* Синтаксические структуры естественного языка в автоматизированных системах общения. М., 1985. 144 с.

Поступила в редколлегию 14.05.90.