УДК 681.324

Д.В.Михайлов, Г.М.Емельянов

ПОСТРОЕНИЕ МОДЕЛИ ОБЪЕКТА ИНФОРМАЦИОННОГО ПРОСТРАНСТВА ПРИМЕНИТЕЛЬНО К ИССЛЕДОВАНИЮ ДИНАМИКИ ФУНКЦИОНИРОВАНИЯ Δ-ГРАММАТИК

The paper deals with the functional and logic structure of information filling in problem of modelling the rule input/output as the tree grammar information space's object.

Как известно, традиционные подходы к формализации преобразований синтаксических структур естественных языков (ЕЯ) в той или иной мере основаны на Δ-грамматиках [1] как формальном аппарате для работы с помеченными деревьями. Тем не менее, ряд задач анализа текстов на ЕЯ наряду с описанием допустимых преобразований деревьев зависимостей требует привлечения знаний о возможных последовательностях таких преобразований с качественным анализом каждой из них. К числу таких задач относятся, в частности, задачи установления смысловой (семантической) эквивалентности ЕЯ-текстов [2], а также решаемая авторами настоящей статьи задача распознавания семантических повторов в сравниваемых по смыслу высказываниях [3]. Решение указанных задач особенно актуально при построении интерпретаторов тестовых заданий открытой формы для систем компьютерного дистанционного обучения.

Для исследования динамики функционирования совокупности правил синонимических преобразований деревьев глубинного синтаксиса [4] при использовании заменяемого лексическим правилом поддерева лексической синонимической конструкции (ЛСК) [3] в качестве элемента повтора в указанной задаче авторами предложена описанная в [5] информационно-логическая модель системы правил расширенной лексико-синтаксической Δ-грамматики. Предложенная модель учитывает недетерминированный характер порождения Δ-грамматикой множества помеченных деревьев, построение целевого вывода сводится к классическим задачам теории сетей Петри.

Однако рассмотрение входа/выхода правила в качестве объекта информационного пространства требует формального описания его активизации как информационного элемента в зависимости от ситуации использования и с учетом его внутренней структуры, для чего необходимо решение двух основных задач:

построение модели входа/выхода правила как объекта информационного пространства;

разработка структуры информационного наполнения анализируемого дерева.

При этом основным требованием к модели входа/выхода правила является отображение различных способов его использования при единообразии функционального описания. Анализ вызывающих активизацию входа/выхода правила событий позволяет выделить следующие способы его использования как информационного элемента:

— анализ применимости правила к помеченному дереву с выдачей FALSE/TRUE в качестве результата;

— синтез дерева по задаваемому выходным деревом шаблону;

— распознавание ключевого слова заменяемого лексическим правилом поддерева;

— расстановка композиционных меток [1] в анализируемом дереве.

Во всех четырех показанных ситуациях элементы информационного пространства активизируются по-разному ввиду неоднородности вызывающих их активизацию событий при идентичности функциональной структуры процессов активизации. Поскольку задача применения правила π к некоторому заданному дереву T_χ есть частный случай задачи «Изоморфизм подграфу», логико-функциональная структура информационного наполнения

входного/выходного дерева T_π правила π должна быть идентична логико-функциональной структуре информационного наполнения анализируемых деревьев.

Действительно, если дерево глубинного синтаксиса фразы χ представить как

$$T_\chi = \langle W_\chi, V_\chi \rangle,$$

где W_χ — множество узлов, V_χ — множество ветвей дерева T_χ, то аналогичным деревом представляется вход/выход T_π правила π:

$$T_\pi = \langle W_\pi, V_\pi \rangle,$$

где элементы множеств W_π и V_π содержат функциональные требования к содержимому узлов и ветвей заменяемого/заменяющего дерева.

С учетом указанных требований авторами предложена структура информационного наполнения узла входного/выходного дерева, унифицируемая со структурой соответствующего описания преобразуемых правилами деревьев и ориентированная на представление динамических структур данных в нотации функционального языка Microsoft muLISP.

В соответствии с приведенным в работах И.А.Мельчука описанием уровня глубинного синтаксиса в информационном наполнении узла глубинной синтаксической структуры следует выделить *лексическую часть*, соответствующую представленному в узле элементу глубинной лексики, и *грамматическую часть*, содержащую семантические словоизменительные характеристики. Кроме того, в описание узла должны быть введены особые элементы, соответствующие пометке входящей в узел ветви и композиционной метке. Исходя из этого, информационное наполнение узла $w_\chi \in W_\chi$ может быть представлено списком из четырех элементов (аналогичным списком представляется информационное наполнение $w_\pi \in W_\pi$):

$$w_\chi = \left(lex\_in_\chi, gram\_in_\chi, arrow\_label_\chi, composition\_label_\chi\right), \quad (1)$$

в котором элемент lex_in_χ соответствует лексической части узла, $gram\_in_\chi$ — грамматической его части, $arrow\_label_\chi$ — пометке входящей ветви, а $composition\_label_\chi$ — композиционной метке узла.

Лексическая часть lex_in_χ узла представляется списочной структурой, первый элемент C_0 которой соответствует самостоятельной лексеме, лексической производной от которой (в виде последовательно взятых значений лексических функций из списка $fun_n,\ldots,fun_1$) является соответствующая содержимому узла лексема (на поверхностно-синтаксическом уровне):

$$lex\_in_\chi = \left(C_0, fun_n,\ldots,fun_1\right),$$

причем список $fun_n,\ldots,fun_1$ может быть пустым в случае отображения в узле фиктивной лексемы, идиомы, либо самостоятельной лексемы, не являющейся лексическим коррелятом [4] в виде значений лексических функций, присутствующих в той же глубинной синтаксической структуре других лексем.

Грамматическая часть узла представляется упорядоченной двойкой:

$$gram\_in_\chi = \left(part\_of\_speech, list\_semant\_categ\right),$$

где $part\_of\_speech$ — символьный атом, обозначающий часть речи, $list\_semant\_categ$ — список семантически обусловленных [4] словоизменительных категорий (у существительных — число, у глаголов — вид, время, наклонение).

Элемент $arrow\_label_\chi$ списочного описания (1) принимает целочисленные значения одного из шести описанных в [1,4] типов связей между родительским и дочерним узлом в

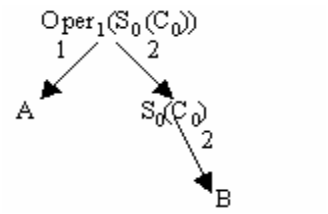
глубинной синтаксической структуре, а для вершины дерева $arrow\_label_{\chi} = 0$ (входящая ветвь отсутствует).

Описание информации узла в виде списка (1) позволяет:

а) формально определить функциональные требования к узлу глубинной синтаксической структуры при описании компонент заменяемого некоторым лексическим правилом дерева; при этом символ C_0 выступает в качестве служебного — им задается местонахождение ключевого слова ЛСК;

б) вычислять значение суперпозиции лексических функций из списка $fun_n,\ldots,fun_1$ с использованием их имен в качестве функциональных аргументов.

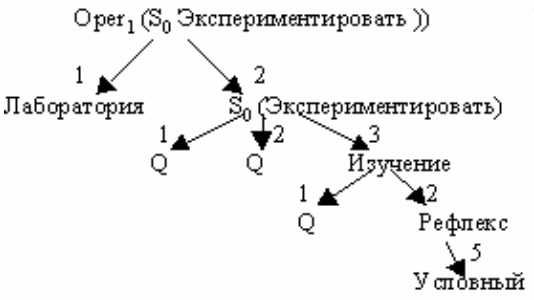
Само дерево представляется составной структурой, первым объектом которой является описание вершины в виде (1), вторым — список дочерних поддеревьев. В нотации Microsoft muLISP такой структуре будет соответствовать список с отсутствием ограничений на число хвостов. Пример подобного описания для входа лексического правила №17 с обслуживающим его синтаксическим правилом №6 при применении в обратном направлении [6] приведен на рис.1. Аналогичный пример приведен на рис.2 для глубинной синтаксической структуры простого распространенного предложения русского языка «Лаборатория провела эксперименты по изучению условных рефлексов».




```
(((C0 S0 Oper1) nil 0 0)
   ((nil nil 1 1))
   (((C0 S0) nil 2 2)
      ((nil nil 2 2))
   )
)
```

а) б)

Рис.1. Входное дерево правила (а) и его списочное описание в нотации Microsoft muLISP (б): A, B — произвольные слова, в процессе перифразирования остаются без изменений




```
(((Экспериментировать (S0 Oper1))
      (V (сов_вид прош_вр изъявит_накл)) 0 nil)
   (((Лаборатория nil)(S (ед_ч)) 1 nil))
   ((Экспериментировать (S0))(S (мн_ч)) 2 nil)
      (((Q nil)(S nil) 1 nil))
      (((Q nil)(S (на)) 2 nil))
      (((Изучение nil)(S (по ед_ч)) 3 nil)
         (((Q nil) nil 1 nil))
         (((Рефлекс nil)(S (мн_ч)) 2 nil)
            ((Условный nil)(A (мн_ч)) 5 nil)
         )
      )
   )
)
```

а) б)

Рис. 2. Дерево глубинного синтаксиса для простого распространенного предложения русского языка (а) и соответствующее ему списочное описание (б)

Поскольку каждое лексическое синонимическое преобразование в общем случае обслуживается одним или несколькими синтаксическими, входное дерево лексического преобразования следует рассматривать как поддерево входного дерева первого из обслуживающих данную лексическую замену синтаксических преобразований. Причем для синтаксических преобразований значимой является только разметка ветвей. Поэтому не относящиеся к описанию ЛСК узлы дерева T_{π} представляются пустым или неопределенным (nil) значением элементов lex_in_{π} и $gram\_in_{\pi}$ списочного описания (1).

При наличии описания T_π и T_χ в виде представленных на рис.1б и рис.2б списочных структур T_π может рассматриваться как система, порождающая отличные друг от друга процессы с идентичной функциональной структурой. Прохождение очередного узла $w_{\pi i} \in W_\pi$ при рекурсивной обработке может быть рассмотрено как абстрактное событие, а установление функционального соответствия некоторого $w_{\chi j} \in W_\chi$ заданному $w_{\pi i}$, размещение в $w_{\chi j} \in W_\chi$ композиционной метки узла $w_{\pi i}$, синтез $w_{\chi k}$ по представленному в $w_{\pi i}$ шаблону — как разные варианты реализации этого события (конкретные действия в процессах анализа применимости правила к помеченному дереву, расстановки композиционных меток в анализируемом дереве и синтеза дерева по заданному выходным деревом правила шаблону). Показанное свойство предложенной модели входа/выхода правила π позволяет оценить ее адекватность с применением методов сетевого моделирования указанных процессов.

Действительно, если прохождению каждого из узлов $w_{\pi i} \in W_\pi$ сопоставить переход $t_i \in T$, а с каждым прохождением узла как разовой реализацией факта изменения некоторого условия связать позицию $p_j \in P$, то работа входа/выхода правила Δ-грамматики моделируется сетью Петри

$$N = \{P, T, F, H, C, M_0\},$$

где P — множество позиций; T — множество переходов; F и H — матрицы инцидентности, $F : P \times T \to \{0,1\}$ и $H : T \times P \to \{0,1\}$; $C = \{color1, color2, color3, color4, color5\}$ — множество цветов маркера; $M_0 : P \to \{0,1\}$ — начальная маркировка или разметка.

Каждому из $\underset{i=1}{\overset{5}{color}}\, i \in C$ соответствует определенный способ использования информационного элемента как вариант разовых реализаций событий прохождения узлов $w_{\pi j} \in W_\pi$ при обходе T_π: $color1$ — анализ применимости правила, $color2$ — синтез дерева на выходе правила, $color3$ — определение ключевого слова ЛСК, $color4$ — расстановка композиционных меток в T_χ.

Следует отметить важные особенности сети N, актуальные для моделирования активизации T_π как объекта информационного пространства с учетом последовательности действий в порождаемых входом/выходом правила π процессах. С целью формального представления окончания обхода дерева T_π как системного события множество переходов T содержит особый переход $t_{out} \Leftrightarrow t_{|T|}$, инцидентный всем $p_i \in P$, для которых $\sum_{j=1}^{|T|} F_{ij} = 0$. Для обозначения изменения условия, соответствующего завершению процесса обхода дерева T_π, множество позиций содержит позицию $p_{out} \Leftrightarrow p_{|P|}$, инцидентную единственному переходу t_{out}. Поскольку в случае успешного завершения анализа применимости правила π к помеченному дереву T_χ последующая перестройка исходного дерева требует идентификации ключевого слова заменяемой ЛСК и расстановки композиционных меток в анализируемом дереве, для задания последовательности указанных процессов в структуру сети N введена дополнительная дуга, соединяющая переход t_{out} с позицией p_1, соответствующей началу процесса обхода дерева T_π. С целью формализации условия окончания анализа/синтеза во избежание развертывания бесконечных процессов в сети N в множество C введен нейтральный маркер $color5$, запрещающий срабатывание перехода, а для перехода t_{out} задается индивидуальная таблица условий срабатывания.

Условия срабатывания перехода t_{out} сети N

$p_i \in P : F[i,\|T\|] = 1,\ i = 1,\ldots,\|P\|$	p_{out}	p_1
$color1$	$color3$	$color3$
$color3$	$color4$	$color4$
$color4$	$color5$	$color5$
$color2$	$color5$	$color5$

Для разрешения конфликтных ситуаций при сетевом моделировании рекурсивной обработки леса дочерних поддеревьев узла $w_{\pi i} \in W_\pi$ в множество переходов T введены безусловные переходы $t_k : \sum_{j=1}^{|P|} H_{kj} > 1$, а прохождение каждого узла представлено двумя позициями сети: до и после прохождения. Пример сетевой модели для представленного на рис.1а входа правила приведен на рис.3.

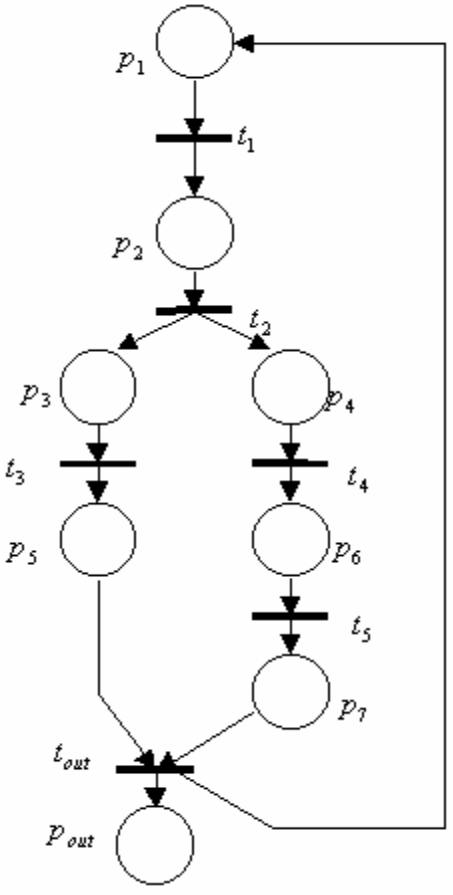


Рис.3. Сетевая модель входа/выхода правила: переход t_1 соответствует прохождению вершины, t_3 — узла A, t_4 — узла с содержимым $S_0(C_0)$, t_5 — узла B

Сеть N обладает рядом свойств, позволяющих оценить адекватность порождаемых ею процессов моделируемым процессам, порождаемым T_π как системой при анализе применимости правила π к дереву T_χ либо синтезом результирующего дерева по задаваемому T_π шаблону.

Теорема 1. Все порождаемые сетью N процессы конечны.

Доказательство следует из конечности (по определению) множеств позиций P и переходов T, а также ограничений, наложенных таблицей на срабатывание перехода t_{out}.

Теорема 2. Сеть N является ограниченной.

Доказательство. Сеть N будет ограниченной, если любое ее место ограничено. Как следует из теоремы 1, $\forall p_j \in P$ может содержать максимум по одному маркеру цвета $\overset{5}{\underset{i=1}{color}}\ i \in C$, максимальное количество маркеров в позиции равно трем (для p_{out}), что и служит доказательством ограниченности N.

Таким образом, сетью N порождаются конечные параллельные процессы без альтернатив и конкуренции. Одновременное появление в позиции p_{out} маркеров цветов $color3$, $color4$ и $color5$ (при анализе применимости правила π) либо маркеров цветов $color2$ и $color5$ (при синтезе дерева по задаваемому деревом T_π шаблону) соответствует завершению указанных процессов. Активизация T_π как объекта информационного пространства может быть формально определена как достижение тупиковой разметки в N при успешном завершении процесса анали-

за/синтеза.

Задание системы правил Δ-грамматики массивами ссылок на описание входов/выходов правил и условий их применимости в виде представленных на рис.1б составных структур позволяет программно реализовывать алгоритмы поиска последовательностей преобразований помеченных деревьев на основе предложенной в [5] информационно-логической модели с применением современных средств логического и функционального программирования..

Работа выполнена в рамках проекта РФФИ №03-01-00055 при поддержке гранта №ТОО-3.3-408 Минобразования РФ, в рамках работ по контракту № И 0675 ФЦП «Интеграция».

1. Гладкий А.В., Мельчук И.А. // Информационные вопросы семиотики, лингвистики и автоматического перевода. Вып.1. М., 1971. С.16-41.
2. Emelyanov G.M., Krechetova T.V., Kurashova E.P. // Pattern Recognition and Image Analysis. 2000. Vol. 10. № 4. P.520-526.
3. Emelyanov G.M., Mikhailov D.V. and Zaitseva E.I. // Pattern Recognition and Image Analysis. 2003. Vol. 13. № 3. P.447-451.
4. Мельчук И.А. Опыт теории лингвистических моделей «Смысл⇔Текст»: Семантика, синтаксис. М.: Школа «Языки русской культуры», 1999. 345 с.
5. Михайлов Д.В., Емельянов Г.М. // Изв. СПбГЭТУ «ЛЭТИ». Сер.: Информатика, управление и компьютерные технологии. Вып. 3. СПб., 2003. С.96-102.
6. Мельчук И.А. Указ. соч. С.154.