

Д.В.Михайлов, Г.М.Емельянов

**ПРЕДСТАВЛЕНИЕ СМЫСЛА В ЗАДАЧЕ УСТАНОВЛЕНИЯ СЕМАНТИЧЕСКОЙ
ЭКВИВАЛЕНТНОСТИ ВЫСКАЗЫВАНИЙ**

In represented work the opportunity of use of the semantic information of a pragmatic level is analyzed at construction of a tree of deep syntax with reference to natural language's statement's sense's comparison's problems.

Ряд задач анализа смысла высказываний на естественном языке (ЕЯ) заключается в их взаимном смысловом сопоставлении на предмет тождественности (эквивалентности). В частном случае некоторый смысл принимается за эталон и ставится задача оценить степень близости смысла высказывания заданному «правильному» смыслу. При формализованном описании смыслов в виде деревьев зависимостей, сетевых либо фреймовых структур мы имеем классическую задачу распознавания образов: принадлежит ли некоторый объект множеству объектов, обладающих заданными свойствами? Наиболее актуальной на сегодняшний день практической задачей, требующей сравнение смысла высказывания с эталоном, является интерпретация тестовых заданий открытой формы [1] в системах автоматизированного тестирования и контроля знаний. Тестовые задания открытой формы требуют от обучаемого формулирования развернутого ответа на поставленный системой вопрос. Одной из основных проблем интерпретации подобных тестов является использование испытуемым синонимов, причем как на уровне отдельных слов, так и их сочетаний. Как показывает опыт, во многих случаях синонимия обусловлена варьированием абстрактной лексики [2], встречающейся в большинстве текстов, в то время как предметная лексика остается без изменений. Пример: *экспериментировать на собаках* — *проводить эксперименты на собаках*. Таким образом, при решении подобных задач необходимо установить семантическую эквивалентность высказывания эталону путем анализа синонимического варьирования абстрактной лексики и оценки возможности сведения формального смыслового представления к эталонному.

Ситуации подобных лексико-синтаксических синонимических замен могут быть описаны правилами синонимического преобразования деревьев глубинных синтаксических структур на основе аппарата стандартных лексических функций (ЛФ) в рамках теоретического подхода к языку как преобразователю «Смысл $\leftrightarrow$ Текст» [2]. При этом в общем случае для каждого дерева в смысловом описании анализируемого высказывания строится множество ЛФ-синонимичных ему деревьев и производится поиск каждого дерева из «эталонного» смыслового описания в каждом из ЛФ-синонимических множеств смыслового описания анализируемого высказывания [3]. Сопоставление смыслов считается успешным, если устанавливается соответствие каждого дерева в «эталонном» смысловом представлении ЛФ-синонимическому множеству деревьев в описании смысла анализируемого высказывания.

Для получения совокупности деревьев глубинного синтаксиса необходимо последовательное выполнение морфологического и синтаксического анализа высказывания. При этом для каждого предложения в тексте строится дерево зависимостей (поверхностная синтаксическая структура) со следующей (ориентировочной) структурой информационного наполнения узла (описание приводится в нотации языка Turbo Prolog):

`surf_struct_node = surf_struct_node(norm_word_form(Word, Base, End), Synt_class, Gram_info),`
где составным объектом `norm_word_form` описывается нормализованная форма слова с выделенной основой `Base` и типичным окончанием `End` из набора признаков словоизменения синтаксического класса `Synt_class`, соответствующая узлу дерева словоформа имеет набор грамматических характеристик `Gram_info`.

При построении дерева глубинного синтаксиса на основе дерева синтаксического подчинения привлекаются знания:

— моделей управления — в соответствии с их описанием по Толково-комбинаторному словарю современного русского языка [2] — для идентификации глубинных синтаксических актантов указанных слов;

— базы данных лексических функций — с целью выявления лексических коррелятов [2] самостоятельных лексем в дереве глубинного синтаксиса.

Для применения описанных в [2] правил синонимического перифразирования информации узла дерева глубинного синтаксиса в соответствии с описанием должна иметь следующую ориентировочную структуру:

$dss_node = dss_node(Lex_info, Gram_info, Arrow_label),$

где Lex_info — лексическая часть информационного наполнения узла; $Gram_info$ — грамматическая часть узла, список семантически обусловленных словоизменительных категорий; $Arrow_label$ — пометка входящей в узел ветви, соответствует одному из описанных в [2] типов глубинных синтаксических связей; $Arrow_label = \{1,2,3,4,5,6\} \cup \{0\}$, пометка 0 соответствует вершине дерева.

Лексическая часть информационного наполнения узла описывается упорядоченной двойкой:

$lex_info = lex_info(C0, Lex_fun),$

где Lex_fun — список лексических функций ключевого слова $C0$. Посредством значений лексических функций из списка Lex_fun представляется связь типа «лексема — лексический коррелят» между самостоятельной лексемой $C0$ и лексемой, информация которой представлена в узле дерева глубинного синтаксиса. Например, для лексемы «проводить» (в контексте «эксперимент») подобное описание будет выглядеть как:

$lex_info(«Экспериментировать», [«S0», «Oper1»]),$

что соответствует описанию значения суперпозиции лексических функций $Oper1(S0$ (Экспериментировать)) для лексемы «проводить» в словарной статье слова «эксперимент» Толково-комбинаторного словаря.

Применяемый авторами [3] алгоритм преобразования дерева синтаксического подчинения в глубинную синтаксическую структуру при разметке ветвей дерева глубинными синтаксическими актантами использует представленную в языковой базе знаний информацию моделей управления. Для корректной работы данного алгоритма необходимо иметь формализованное описание моделей управления всех лексем используемого подмножества ЕЯ, с которыми могут находиться в отношении подчинения другие лексемы. Данное требование относится как к абстрактной универсальной, так и к предметной лексике. Описания моделей управления абстрактных лексем могут быть заложены в базу изначально, при построении системы. Предметная лексика описывается в процессе настройки на конкретную область знаний, а также в процессе эксплуатации системы. Построение моделей управления для новых слов предполагает наличие у сопровождающих систему специалистов лингвистических навыков. Однако в лучшем случае квалифицированный пользователь (настройщик) обладает навыками описания предметной области в некоторой известной формальной нотации и реализации ее на компьютере средствами предлагаемого программой визуального конструктора (т.е. может в буквальном смысле «нарисовать» на экране объекты, атрибуты и отношения). Поэтому актуальной здесь является проблема автоматизации пополнения языковой базы знаний информацией моделей управления предметных лексем.

Рассмотрим возможность решения указанной проблемы за счет привлечения знаний предметной области в формальной нотации иерархии семантических классов заданного регистра (жанра) языка и сетевого описания модели предметной области [4,5]. При использовании, в частности, формального концептуального анализа [6] подобные описания могут быть получены автоматически, по введенным настройщиком текстам.

Отношение между лексемой и ее глубинными синтаксическими актантами по модели управления может быть представлено следующей структурой:

$t_government_pattern(Lexema, t_surf_real(Prep, Synt_class, Gram_info),$
 $Arrow_label,$
 $t_semantic_contence(Role_name, Sem_class)).$

Здесь составным объектом `t_surf_real(Prep, Synt_class, Gram_info)` описывается способ поверхностной реализации соответствующего актанта лексемы `Lexema` с учетом предлога (`Prep`), который должен присутствовать в поверхностной реализации (поле может быть пустым). Элемент `Argow_label` соответствует типу отношения подчинения (в реальных моделях управления это 1, 2, 3 либо 4). Структурой `t_semantic_contence` описывается семантическая интерпретация глубинного синтаксического актанта: `Role_name` — название роли обозначаемой актантом сущности, относящейся к семантическому классу `Sem_class`.

Следует отметить, что теоретическое описание моделей управления лексем в Толково-комбинаторном словаре имеет ряд сходств с описанием концептуальных структур (концептуализаций) при использовании последних в семантически-ориентированном подходе к анализу ЕЯ.

Во-первых, глубинному синтаксическому актанту лексемы C_0 ставится в соответствие его семантическая интерпретация. Семантическая интерпретация глубинного синтаксического актанта слова C_0 в его модели управления есть роль обозначаемой этим актантом сущности в обозначаемой C_0 ситуации. При этом каждой из предопределенных [4] ролей может быть приписана глубинная синтаксическая характеристика в соответствии с приведенной И.А.Мельчуком классификацией отношений глубинного синтаксиса:

- роли `who` (агенса) — тип подчинения 1,
- роли `what` (пациенса) — тип подчинения 2,
- роли обстоятельства места `Where` — тип подчинения 3,
- роли обстоятельства времени `When` как наиболее удаленного от главного слова-концепта во фразе — тип подчинения 4.

Во-вторых, каждой семантической интерпретации можно поставить в соответствие некоторый семантический класс обозначаемой актантом сущности. Эти семантические классы выражают сходный смысл лексических единиц в разных языковых жанрах. Так, для приведенного в [2] примера модели управления глагола «колотить» семантической ориентации первого глубинного синтаксического актанта (агенса) будут соответствовать семантические классы «существо» либо «отдельный и цельный предмет (не являющийся частью другого предмета и не имеющий отчетливых частей)». Сами по себе семантические классы образуют иерархию, причем некоторые из них могут быть общими для разных жанров:

`d_sem_class(t_sem_class(Sem_class, Parent_sem_class, Sem_orientation))`.

Здесь `Parent_sem_class` — семантический класс, родительский для `Sem_class`; `Sem_orientation` — ориентация семантического класса `Sem_class` на определенный элемент модели предметной области. В приведенном примере оба семантических класса могут быть отнесены к более абстрактному классу «объект».

Показанные свойства модели управления позволяют:

- использовать информацию концептуальных структур при установлении отношения глубинного синтаксического подчинения между главными (концептуальными) и зависимыми (ролевыми) предметными лексемами;
- использовать информацию системы семантических классов для более точной идентификации глубинных синтаксических актантов лексем в соответствии с требованиями их моделей управления.

Если семантические классы (`Sem_class`) и ориентация (`Sem_orientation`) предметных лексем на элементы модели предметной области (классы, типы, атрибуты [4]), а также ролевая ориентация (если такая имеется) указаны в словаре предметной лексики:

`d_subj_lexicon(t_subj_lexicon(Lexema, Sem_class, Sem_orientation, Role_name))`,

то информацию иерархии семантических классов можно использовать при установлении отношения глубинного синтаксического подчинения между некоторой заданной лексемой `Lexema` (не обязательно абстрактной) и предметной лексемой `Subj_lexema` на основе имеющегося в базе данных описания модели управления (ролевая ориентация предметной лексики здесь не учитывается, о ней ниже):

```

deep_syntax_relation(dss_node(lex_info(C0, Lex_fun), _, _),
    dss_node(lex_info(Subj_lexema, []), Gram_info_subj_lex, Arrow_label)) :-
lex_fun_list_make(Lexema, C0, Lex_fun),
d_government_pattern(t_government_pattern(Lexema,
    t_surf_real(_, Synt_class_subj_lex, Gram_info_subj_lex),
    Arrow_label,
    t_semantic_contence(_, Sem_class))),
d_subj_lexicon(t_subj_lexicon(Lexema, Sem_class_subj_lex, _, _)),
correspondence(Sem_class_subj_lex, Sem_class),
morfol_characteristics(Lexema, Synt_class_subj_lex, Gram_info_subj_lex).

```

Здесь: `lex_fun_list_make` — Пролог-правило построения списка `Lex_fun` символов лексических функций для лексемы `Lexema`, являющейся значением суперпозиции лексических функций некоторой другой самостоятельной лексемы `C0`; `correspondence` — Пролог-правило установления соответствия семантического класса `Sem_class_subj_lex` предметной лексемы семантическому классу `Sem_class` по модели управления в соответствии с имеющейся иерархией семантических классов; `morfol_characteristics` — Пролог-правило выявления морфологических характеристик (в экспериментальном варианте был реализован морфологический анализ на основе словаря основ по описанному в [7] алгоритму). Построение списка символов лексических функций можно организовать как рекурсивный просмотр занесенного в динамическую базу данных словаря лексических функций (`d_lf`):

```

lex_fun_list_make(Lexema, C0, [LF | LF_list]) :-
    d_lf(t_lf(Lexema, Lexema1, LF)), lex_fun_list_make(Lexema1, C0, LF_list).
lex_fun_list_make(Lexema, C0, [LF]) :- d_lf(t_lf(Lexema, C0, LF)), not(d_lf(t_lf(C0, _, _))).

```

Значение лексических функций описывается составным объектом `t_lf`, первый атрибут которого соответствует значению лексической функции, второй — аргументу, третий — символному обозначению лексической функции.

С учетом указанного выше соответствия предопределенных ролей глубинным синтаксическим актантам концептуальные структуры могут быть использованы для идентификации отношений глубинного синтаксиса между предметными лексемами. Однако для построения соответствующего фрагмента дерева глубинного синтаксиса информации о ролевой ориентации лексем недостаточно. Требуется информация о способе поверхностной реализации глубинных синтаксических актантов для лексем предметной лексики, обозначающих концепты.

Наличие в словаре предметной лексики ролевых ориентаций предполагает хранение всех словоформ каждой предметной лексемы, значимых для ролевой идентификации. Следовательно, для получения информации о возможных вариантах поверхностной реализации глубинных синтаксических актантов лексемы `Concept`, являющейся именем концептуализации, необходимо выполнить морфологический анализ содержащихся в словаре предметных лексем с ориентациями на роли этой концептуализации.

Будем считать, что концептуализации представляются в динамической базе данных концептуальных структур (`d_conceptualization`) составными объектами вида:

```

t_conceptualization(Concept, t_role(Role_name, Sem_orientation, Arrow_label)),

```

где каждая роль описана составным объектом `t_role` и имеет в качестве атрибутов задаваемое при настройке название роли `Role_name` (`Who`, `What`, `Where`, `When` (см. выше)), семантическую ориентацию `Sem_orientation` на некоторый элемент модели предметной области и заданную при настройке глубинную синтаксическую характеристику `Arrow_label`.

С учетом вышесказанного получаем следующий (ориентировочный) вариант Пролог-правила для построения и добавления к динамической базе данных моделей управления для предметных лексем, обозначающих концептуализации:

```
add_govern_pattern : -  
  d_conceptualization(t_conceptualization(Lexema,  
                                           t_role(Role_name, Sem_orientation, Arrow_label))),  
  d_subj_lexicon(t_subj_lexicon(Subj_lexema, Sem_class, Sem_orientation, Role_name)),  
  morfol_characteristics(Subj_lexema, Synt_class, Gram_info),  
  assertz(d_government_pattern  
          (t_government_pattern(Lexema,  
                                t_surf_real(“”, Synt_class, Gram_info),  
                                Arrow_label,  
                                t_semantic_contence(Role_name, Sem_class)))).
```

В целях наглядности здесь не рассматривается наличие предлога (Prep), обязательно-го в поверхностной реализации.

Использование семантической информации предметной области предложенным способом позволяет строить деревья глубинного синтаксиса для текстов произвольной тематики. При этом недостающая в языковой базе знания информация моделей управления может быть получена на основе анализа концептуальных структур, что особенно актуально в связи с ориентацией оригинального Толково-комбинаторного словаря современного русского языка на абстрактную (универсальную) лексику.

Работа выполнена при поддержке РФФИ (проект №03-01-00055) в рамках работ по контракту № И 0675 ФЦП «Интеграция».

1. Васильев В.И., Демидов А.Н., Малышев Н.Г., Тягунова Т.Н. Методологические правила конструирования компьютерных тестов. М.: МГУП, 2000. 64 с.
2. Мельчук И.А. Опыт теории лингвистических моделей «Смысл $\leftrightarrow$ Текст». Семантика, синтаксис. М.: Языки русской культуры, 1999. 345 с.
3. Emelyanov G.M., Krechetova T.V., Kurashova E.P. // Pattern Recognition and Image Analysis. 2000. V.10. №4. P.520-526.
4. Жигалов В.А. // Тез. докл. Всерос. межвуз. науч.-техн. конф. студентов и аспирантов «Микроэлектроника и информатика-99». Москва, 1999. С.184.
5. Нариньяни А.С. // Тр. Междунар. семинара «Диалог-97» по компьютерной лингвистике и ее приложениям. Москва, 1997. С.203-208.
6. Ganter B. and Wille R. Formal Concept Analysis — Mathematical Foundations. Berlin: Springer-Verlag, 1999. 284 p.
7. Белоногов Г.Г., Новоселов А.П. Автоматизированные информационные системы. М.: Наука, 1973. 328 с.