Управление программной на естественном языке

Е. А. Черкашин

ИДСТУ им. В. М. Матросова СО РАН

7 ноября 2019, Иркутск

Актуальность проблемной области

Направление ИИ – Понимание естественного языка (ЕЯ), перевод из одного ЕЯ на другой. Решаются следующие задачи:

- 1. Анализ текстов, помещение изъятой информации в базу данных:
 - изготовление шаблонов документов, отчетов;
 - синтез структур данных для ИС;
 - заполнение баз данных ИС и т.п.;
- 2. Ведение диалога с пользователем:
 - идентификация моделей и планирование действий (интеллектные пользовательские интерфейсы);
 - приобретение знаний (оболочки экспертных систем);
- 3. Управление приложением:
 - запросы на естественном языке к базам данных;
 - внесение изменений в данные;
 - ▶ командное управление («Проветрить квартиру»).

Графический пользовательский интерфейс

Специализирован на унарных операциях над отдельными частями информационного объекта.





В операциях аргументы вводится в диалоговом окне (для операций с аргументами).

Более «умные» редакторы не используют контекстные операции (EMACS, VI, Visual Studio Code, Sublime, AutoCAD).

Alt-X replace-string, Набирается как «Alt-X repl str»

```
\begin{frame}
\frametitle{---}

%k 100:1198U0---/-/text/pres/iglu/talk-iglu-2019-03-14.tex | 30% /PS Helm + C6 wap | 9 14:33 4.43

Replace string (default \includegr → % \includegr): | includegr → % \includegr
```

Требования к методике

Чего бы мы хотели от такой системы?

- Представление запросов в терминологии предметной области ИС;
- 2. Выполнение двуаргументных операций над объектами целого класса (не только с экземплярами);
- 3. Система должна синтезировать ответ/действие на основе
 - введенного запроса/команды на ЕЯ,
 - данных и знаний предметной области;
- 4. Расширяемость/замена предметной области;
- 5. (в идеале) задавать утончающие вопросы;
- 6. (в идеале) обучаться (синонимы, новые декларативные конструкции).

Пакеты прикладных программ



input a.txt multipy by b.txt output c.txt

Язык

Семиотика (наука о знаках) делится на три раздела (Моррис)

- □ Семантика отношение знака к объекту (естествознание).
- □ **Синтаксис** отношение знаков между собой (лингвистика, логика).
- □ Прагматика отношение знака к субъекту (психология).

Язык программирования

Синтаксис на уровне грамматики определяет корректные последовательности символов (лексемы, операторы, структуры). Но синтаксическая правильность не гарантирует даже осмысленности программы.

Семантика — это соответствие между синтаксически правильными программами и действиями абстрактного исполнителя, то есть это смысл синтаксических конструкций.

Цель программиста — получить нужный ему эффект в результате исполнения программы на конкретном оборудовании: трансляция и исполнение осуществляется на конкретных вычислителях.

Прагматика — задает конкретизацию абстрактного вычислителя для данной вычислительной системы. Стандарт языка программирования не полностью задаёт поведение исполнителя, которое производителями транслятора языка реализуется на конкретной программно-аппаратной платформе.

Реализованный язык является прагматическим опосредованием абстрактной модели вычислений и возможностью ее реализации.

Синтаксический разбор предложения



Грамматика

$$G = \langle T, N, \Sigma, R \rangle$$

Т - множество терминальных символов (слова, буквы, IF, ELSE), N – множество нетерминальных символов (обозначения, A, B, <noun>, $\langle verb \rangle$, $T \cap N = \emptyset$, Σ – стартовый символ (<программа>, <предложение>), $\Sigma \in N$, R – множество правил грамматики $A \rightarrow B$

 $R \subset ((T \cup N)^*N(T \cup N)^*) \times (T \cup N)^*.$

Язык $L(G) = \{\Omega \in T^* | \Sigma \to^* \Omega \}.$

Вывод $\Sigma \to^* \Omega$ $\Sigma \to \Sigma A \quad \Sigma \to A$ $A \rightarrow b\Sigma e$ $A \rightarrow be$

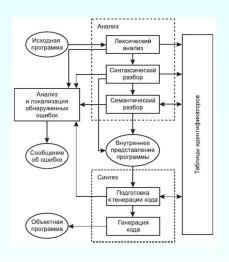
Пример:
$$a = (', b = ')'$$
.

$$\begin{array}{lll} \mathbf{A} & \mathbf{A} \\ b \Sigma e & (\Sigma) \\ b \Sigma \mathbf{A} e & (\Sigma \mathbf{A}) \\ b \mathbf{A} \mathbf{A} e & (\mathbf{A} \mathbf{A}) \\ b b e \mathbf{A} e & (()(\mathbf{A}) \\ b b e b e e & (()()()) \end{array}$$

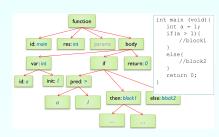
Грамматика языка программирования С

```
<translation-unit> ::= {<external-declaration>}*
<external-declaration> ::= <function-definition>
                  | <declaration>
<function-definition> ::= {<declaration-specifier>}* <declarator> {<declaration>}* <compound-statement>
                  | union
<struct-declaration> ::= {<specifier-qualifier>}* <struct-declarator-list>
<specifier-qualifier> ::= <type-specifier>
                  | <type-qualifier>
<struct-declarator-list> ::= <struct-declarator>
                  <struct-declarator-list> , <struct-declarator>
<selection-statement> ::= if ( <expression> ) <statement>
                  | if ( <expression> ) <statement> else <statement>
                  | switch ( <expression> ) <statement>
<iteration-statement> ::= while ( <expression> ) <statement>
                  | do <statement> while ( <expression> ) :
                  for ( {<expression>}? ; {<expression>}? ; {<expression>}? ) <statement>
<jump-statement> ::= goto <identifier> ;
                    continue ;
                    return {<expression>}?;
```

Транслятор языка программирования



Внутреннее представление программы – абстрактное синтексическое дерево.



Типы грамматик

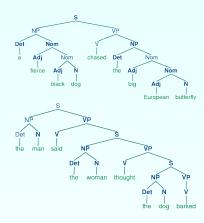
По иерархии Ноама Хомского, грамматики делятся на **четыре** типа, каждый последующий является более ограниченным подмножеством предыдущего (но и легче поддающимся анализу):

- □ Тип о. Неограниченные грамматики возможны любые правила;
- Тип 1. Контекстно-зависимые грамматики левая часть может содержать один нетерминал, окруженный «контекстом»; сам нетерминал заменяется непустой последовательностью символов в правой части;
- □ Тип 2. Контекстно-свободные грамматики левая часть состоит из одного нетерминала;
- □ Тип 3. Регулярные грамматики более простые, распознаются конечными автоматами.

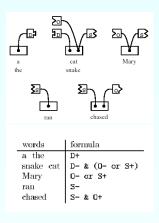
Пример трансляции

```
.file
                "fact.c"
        .globl
                fact
                 fact, Ofunction
        .type
fact:
.LFB11:
        movl
                $1, %eax
                $1. %rdi
        cmpa
        ibe
                 .L4
        imulq
                %rdi, %rax
        suba
                $1. %rdi
                $1, %rdi
        cmpq
        jne
               .L3
.L4:
        ret
        .section
                     .rodata.str1.1, "aMS", @progbits,1
.LCo:
        .string
                     "Factorial of %lu = %lu.\n"
        .section
                     .text.startup."ax".@progbits
        .globl main
                main. Ofunction
        .tvpe
main:
.LFB12:
        ret
        .cfi endproc
        .size
                main. .-main
        .ident "GCC: (GNU) 8.2.1 20181127"
        .section
                     .note.GNU-stack,"", @progbits
```

Трансляция предложений естественного языка



Linked grammar (грамматики связей)



Последовательность слов находится в linked grammar, если существует способ нарисовать связи между словами, такие что

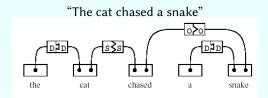
Связи не пересекаются (планарный граф);

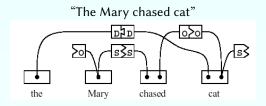
Все слова последовательности соединены связями (связность);

Все связи удовлетворяют ограничениям (непротиворечивость)

Два слова соединены одной и только одной связью (**исключительность**).

Примеры разбора Linked grammar





Примеры разбора Linked grammar

"A dog arrived with a bone"
"A dog with a bone arrived"

Примеры разбора Linked grammar

```
es to insulin-induced hypoglycemia.
                                   0.47 seconds (0.47 total)
++++Time
Found 129 linkages (75 had no P.P. violations)
 Linkage 1. cost vector = (UNUSED=0 DIS=0 AND=0 LEN=39)
              +-----MVp------
              +-----
                  +-----D*u----+
                        --D"U-----+ +--JS--+
+----A---+--Mp-+JS+ | +-DG+
   +---Wd--+---SS--+
LEFT-WALL GABA mediates v the inhibitory a effect n of NO on the AVP and OXT
         +-----+
responses.n to insulin-induced hypoglycemia[?].n .
              +-----MVp------
              +-----
                        +----A---+--Mp-+Js+
   +---Wd--+---Ss--+
LEFT-WALL GABA mediates.v the inhibitory.a effect.n of NO on the AVP and OXT
         +-----+
-AN--+
               +-----Ah-----+
responses.n to insulin-induced hypoglycemia[?].n .
```

Информационная система GeoBase. База данных

GeoBase – программа, позволяющая делать запросы на ЕЯ к базе данных по географии США, Borland, 1988.

```
state('alabama', 'al', 'montgomery', 3894e3,51.7e3,22, 'birmingham', 'mobile', 'montgomery', 'huntsville').
state('alaska'.'ak'.'iuneau'.401.8e3.591e3.49.'anchorage'.'fairbanks'.'iuneau'.'sitka').
city('alabama','al','birmingham',284413).
city('alabama'.'al'.'mobile'.200452).
border('florida'.'fl'.['georgia'.'alabama']).
highlow('alabama','al','cheaha mountain',734,'gulf of mexico',0).
mountain('alaska'.'ak'.'mckinlev'.6194).
mountain('alaska', 'ak', 'st. elias', 5489).
road('66',['district of columbia','virginia']).
lake('huron'.59570.['michigan']).
  The database contains the following information:
Information about states:
  Area of the state in square kilometers
  Population of the state in citizens
  Capital of the state
  Which states border a given state
  Rivers in the state
  Cities in the state
  Highest and lowest point in the state in meters
Information about rivers:
  Length of river in kilometers
Information about cities:
  Population of the city in citizens
```

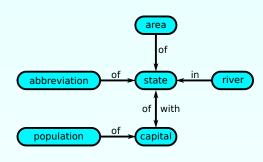
Примеры запросов

Some sample queries:

- states
- give me the cities in california.
- what is the biggest city in california ?
- what is the longest river in the usa?
- which rivers are longer than 1 thousand kilometers?
- what is the name of the state with the lowest point?
- which states border alabama?
- which rivers do not run through texas?
- which rivers run through states that border the state with the capital austin?

Схемы (спецификации) интерпретации

```
schema('abbreviation','of','state').
schema('state','with','abbreviation').
schema('capital'.'of'.'state').
schema('state','with','capital').
schema('population'.'of'.'state').
schema('area','of','state').
schema('city','in','state').
schema('length','of','river').
schema('state','with','river').
schema('river','in','state').
schema('capital'.'with'.'population').
schema('point','in','state').
schema('height','of','point').
schema('mountain','in','state').
schema('height','of','mountain').
schema('lake'.'in'.'state').
schema('name','of','river').
schema('name'.'of'.'capital').
schema('road','in','continent').
```



Программа

Основной цикл программы

Синтаксический анализатор (часть)

population of Washington Население штата или города? 1? schema('population','of','city'). 2? schema('population','of','state').

Корпус реализован при помощи реструктуризации базы данных.

```
% . . . . . . . . . . . . . . . /* Relationships about states */
db(abbreviation.of,state,ABBREVIATION,STATE):- state(STATE,ABBREVIATION,,,,,,,).
db(state,with,abbreviation,STATE,ABBREVIATION):-state(STATE,ABBREVIATION,_,_,,,,,,,,).
db(area.of.state.AREA.STATE):-
                                      state(STATE, , , , AREA1, , , , , ), str real(AREA, AREA1).
db(capital,of,state,CAPITAL,STATE):- state(STATE,_,CAPITAL,_,_,_,_).
db(state,with,capital,STATE,CAPITAL):-state(STATE,_,CAPITAL,_,_,_,_,_).
db(population,of,state,POPULATION,STATE):-state(STATE,,,POPUL,,,,,,,),str real(POPULATION,POPUL).
db(state,border,state,STATE1,STATE2):-border(STATE2,_,LIST),member(STATE1,LIST).
/* Relationships about rivers */
db(length.of.river.LENGTH.RIVER):-
                                      river(RIVER.LENGTH1. ).str real(LENGTH.LENGTH1).
db(state,with,river,STATE,RIVER):-
                                      river(RIVER, ,LIST), member(STATE,LIST).
db(river,in,state,RIVER,STATE):-
                                      river(RIVER,_,LIST), member(STATE,LIST).
/* Relationships about points */
db(point,in,state,POINT,STATE):-
                                      highlow(STATE,_,POINT,_,_,_).
db(point.in.state.POINT.STATE):-
                                      highlow(STATE. . . . POINT. ).
db(state,with,point,STATE,POINT):-
                                      highlow(STATE,_,POINT,_,_,_).
db(state.with.point.STATE.POINT):-
                                      highlow(STATE,_,_,POINT,_).
db(height, of, point, HEIGHT, POINT):-
                                      highlow( , , , , POINT, H), str real(HEIGHT, H),!.
db(height, of, point, HEIGHT, POINT):-
                                      highlow(_,_,POINT,H,_,_),str real(HEIGHT,H),!.
```

Программа

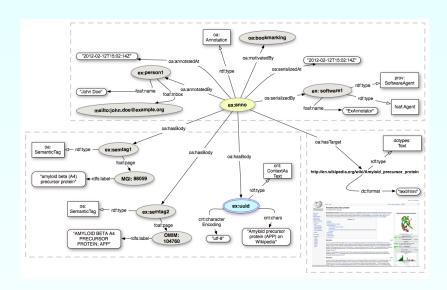
Интерпретация запроса ...findall(A, eval_interp(Q, A), L),

```
•••
  eval interp(0.IAns):-
                eval(Q,A),
                e i(A.IAns).
  eval(q min(ENT.TREE).ANS):-
                findall(X.eval(TREE.X).L).
                entitysize(ENT,ATTR),
                sel_min(ENT,ATTR,99e99,''.ANS,L).
  eval(q max(ENT,TREE),ANS):-
                findall(X,eval(TREE,X).L).
                entitysize(ENT,ATTR),
                sel max(ENT.ATTR.-1.''.ANS.L).
  eval(q sel(E,gt,ATTR,VAL),ANS):-
                schema(ATTR.ASSOC.E).
                db(ATTR, ASSOC, E, SVAL2, ANS),
                str real(SVAL2.VAL2).
                VAL2>VAL.
%
eval(q_eaq(E1,A,E2,TREE),ANS):-
                eval(TREE.VAL).db(E1.A.E2.ANS.VAL).
  eval(g eaec(E1.A.E2.C).ANS):-db(E1.A.E2.ANS.C).
  eval(q e(E),ANS):-
                            ent(E,ANS). % EVAL "ATOM"
  eval(q_or(TREE,_),ANS):- eval(TREE,ANS).
  eval(q_or(_,TREE),ANS):- eval(TREE,ANS).
  eval(q and(T1.T2).ANS):- eval(T1.ANS1).eval(T2.ANS).ANS=ANS1.
```

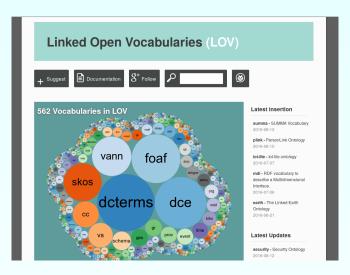
Семантический WEB

В настоящее время появились новые варианты представления предметной области, при этом
предметной области, при этом
🗅 сущности идентифицируются, как правило, глобально;
🗅 разработаны стандарты представления информации в
конкретных предметных областях (oa, prov, foaf,);
🗅 стандартизированы базовые онтологии (RDF, RDFs, dc, dcterms),
при помощи которых задаются предметные онтологии;
исследуются свойства логик описаний (DL, SHOIND,);
предложены способы;
□ созданы форматы представления данных (RDF, RDFa, OWL, JSON-LD,);
□ существуют сервисы предоставления онтологической
информации (dbpedia.org, lov.linkeddata.es/dataset/lov/),
программное обеспечение создания таких сервисов (ClioPatria);
□ язык запросов к онтологическим сервисам (SPARQL), технологии
распределенных запросов к сервисам;
□ инструментальные средства моделирования предметной области
с выдачей в стандартные форматы (Protegé).

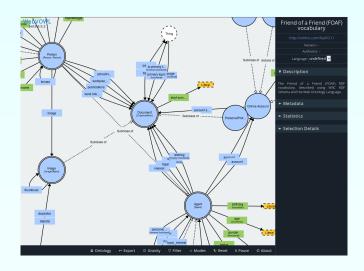
Онтология Open Annotation (oa)



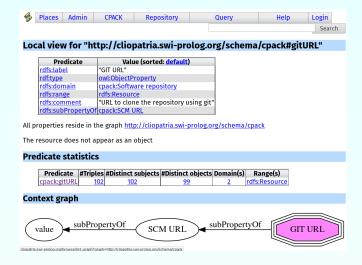
Инструменты: словарь связных данных (LOV)



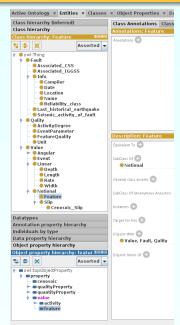
Инструменты: браузер онтологий



Инструменты: сервер онтологий ClioPatria



Инструменты: редактор онтологий Protegé

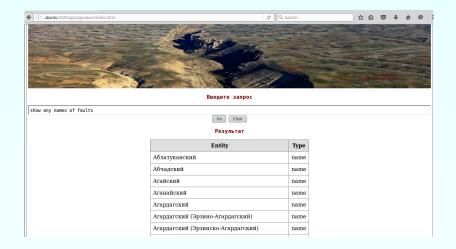


```
eugeneai@center datal$ head -n 50 activity fall data.ttl
prefix foaf: <http://xmlns.com/foaf/0.1/>
prefix geob: <a href="mailto://www.semanticweb.org/bernard">prefix geob: <a href="mailto://www.semanticweb.org/bernard">http://www.semanticweb.org/bernard</a> black/ontologies/2016/3,
@prefix nie: <http://www.semanticdesktop.org/ontologies/2007/01/19/nie#>
dprefix owl: <http://www.w3.org/2002/07/owl#>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
prefix xml: <http://www.w3.org/XML/1998/namespace>
prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
[] a geob:Fault ;
   nie:identifier "RUAF 235-1"
   nie:title "Северобайкальский" :
   geob:activity [ a geob:Activity :
            geob:degree "повышенная" ;
            geob:eventage [ a geob:EventAge ;
                     geob:index 2e+00 ;
                     geob:type geob:LastActivationAge :
                     geob:value "Голоценовое" ] :
            geob:value 1.4e+01 ] ;
   geob:angle [ a geob:Angle ;
            geob:quality "ЛС" ;
            geob:value "50-80" 1 :
   geob:azimuth [ a geob:Azimuth ;
            geob:quality "ЛС" ;
            geob:value 1.02e+02 ] ;
   geob:cenosoicslip [ a geob:CenosoicSlip :
            geob:guality "ЛС" :
            geob:reliabilityClass 1e+00 ;
            geob:type geob:vertical;
            geob:value 4e+03 ] :
   geob:compiler [ a geob:Compiler :
            nie:created "15.11.2014" :
            foaf:name "Лунина О.В." ];
   geob:event [ a geob:Event ;
            qeob:associatedCSS "Северобайкальский" :
            geob:averaged_slip_rate_mm_year 3e+00 ;
            geob:isActiveFault "Да"
            geob:potential ms max 7.7e+00 ;
            geob:potential mw max 0e+00 :
            geob:guality "ЛС":
            geob:slip rate mm year "1-4.99" :
            geob:type geob:TotalMaxSlip ;
            geob:value 9e+00 ].
        Γ a geob:Event :
            geob:type geob:LateralMaxSlip ;
            geob:value 0e+00 ],
        [ a geob:Event ;
            geob:type geob:VerticalMaxSlip ;
            geob:value 9e+00 1 :
     eob:feature [ a geob:Feature
```

Модификация GeoBase на основе Semantic WEB

```
schema('fault','in','continent'). % Связь с терминами GeoBase
schema('fault','with','feature').
schema('name','of','fault').
% schema('feature','of','fault'). % Это отношение находится в T-Box Faults''
% «Загрузка» онтологий
schema(Prop, 'of', SubjName):-
                                  % используется на этапе трансляции
        var(SubjName),
        geob prop(Prop. ).
schema(Prop, 'of', SubjName):-
                                  % используется на этапе интерпретации
        nonvar(SubiName).
                                  % задан конкретный класс
        geob prop(Prop, GProp),
        geob_ent_class(SubjName, Subj),
        rdf reachable(Subi. rdfs:subClassOf. Parent).
        rdf(GProp, rdfs:domain, Parent),!.
geob prop(Prop. GProp):- % Проверка свойства
        rdf global id(geob:Prop, GProp),
        rdf(GProp.rdf:type.owl:'ObjectProperty').!.
geob class(Class, GClass):- % Проверка класса
        rdf global id(geob:Class. GClass).
        rdf(GClass,rdf:type,owl:'Class'),!.
geob ent class(Ent. Class):-
        sub atom(Ent,0,1,R,H),
        sub atom(Ent.1.R.0.T).
        string upper(H,U),
        atom concat(U,T,GEnt),
        geob class(GEnt. Class).
geob_ent(E, A):-
        nonvar(E).
        geob ent class(E,Class),
        rdf(A. rdf:tvpe. Class. geodata). % Сущность онтологии geodata
```

Интерфейс пользователя GeoBase + ActiveFaults



Схемы данных из моделей ИС

В разработке программного обеспечения используются модели, представляющие объекты, бизнес-процессы и организационную систему предприятия (UML-2.4, SysML, BPMN-2.0, CMMN).

- 1. Формирование модели при помощи редакторов;
- 2. Экспорт модели в формат ХМІ;
- Преобразование XMI в формат онтологии <subject, predicate, object>;
- 4. Преобразование базы данных в формат онтологии, или реализация адаптера;
- 5. Настройка связей с основной схемой данных;
- 6. Настройка корпуса, который также можно организовать при помощи адаптера базы данных.

Спасибо за внимание!