ОПЫТ ПОСТРОЕНИЯ ПРЕДИКАТНЫХ ФОРМ ПРЕДЛОЖЕНИЙ
AN ATTEMPT AT THE CONSTRUCTION OF PREDICATIVE FORMS FOR SENTENCES

Г.В.Лезин, С.М. Герасимов, Е.А. Каневский
Санкт-Петербургский экономико-математический институт РАН
lezin@emi.nw.ru

Рассматриваются представление предикатной формы предложения, преобразование семантико-синтаксической модели в предикатную форму и общий алгоритм вычисления предиката в контексте предложения. В качестве исходных взяты семантический словарь и семантико-синтаксическая модель предложения, разработанные В.А. Тузовым.

1. Введение

Можно, по-видимому, считать устоявшимся взгляд на предложение естественного языка как на сообщение, содержащее три вида информации [1–3]:

а) Внеконтекстная информация, опирающаяся исключительно на смысл используемых в предложении слов. Информация этого вида может быть представлена в виде пропозициональной формы (предиката), переменные которого сопоставляются лексическим значениям слов предложения (лексемам) и определены на множествах внеязыковых сущностей, обозначаемых этими словами в конкретных контекстах.

б) Контекстная информация образуется в результате вычисления предиката в контексте предложения. В случае компьютерного анализа текста контекст образован базой знаний о предметной области текста и информацией из текста, предшествующего этому предложению. Вычисляя предикат предложения, мы определяем: является ли значение переменной новым, отсутствующим в контексте, либо известным.

в) Иллокутивная функция предложения. Информация этого вида отражает цели и намерения говорящего, его отношение к высказываемому.

Исходными для построения и вычисления предиката являются разработанные В.А. Тузовым семантический словарь русского языка и семантико-синтаксическая модель предложения [4, 5]. Предложение представлено в модели деревом подчинительных связей между его лексемами. В семантическом словаре каждой из лексем дано толкование, записанное на специально разработанном для этой цели формальном языке.

Были разработаны программы трансляции толкований на язык системы управления базой знаний MAZE [6] и проведены эксперименты по вычислению предикатов предложений в контексте базы знаний. Анализ результатов проделанной работы показал необходимость уточнения формальной интерпретации основных конструкций языка и синтаксических связей в семантико-синтаксических моделях предложений.

Статья посвящена формализации и уточнению денотативного статуса лексем языка, а также способам его отображения в словаре и учета в алгоритмах построения и вычисления предикатов предложений.

2. Семантический язык В.А. Тузова

2.1. Основой описания формальной семантики лексем в семантическом словаре является классификатор основных понятий русского языка. Классификатор имеет иерархическую структуру с отношением "один" ← "многие". Обозначение класса снабжено префиксом '$'.

Пример.

Лексеме ВРУЧЕНИЕ$15210 соответствует следующее место в иерархии классов: "существительное" (класс $1) ← "действие" ($15) ← "занятие" ($1521) ← "приобретение" ($15210).

По принципу классификации лексемы поделены на две категории: объекты и их свойства. Каждому из классов может быть сопоставлена собственная иерархия свойств. Если лексема L, относясь к классу C, трактуется как его свойство из класса P, то класс лексемы L обозначается в словаре конструкцией L$C/P

Пример.

Лексемы СТАРЫЙ$12411/071 и ЮНЫЙ$12411/071 принадлежат общему классу "существительное" ($1) ← "физический объект" ($12) ← "живой"($124) ← "человек" ($1241) ← "индивид"($12411) и относятся к общему классу его (класса $12411) свойств - "возраст" (/07) ← "молодой-старый" (/071).

Лексему в словаре мы также будем трактовать как класс, объединяющий (потенциально) полное множество внеязыковых сущностей (денотатов), обозначаемых этой лексемой во всевозможных контекстных ситуациях. Бинарные базисные функции мы интерпретируем как классы бинарных отношений. Элементами этих классов являются бинарные отношения, связывающие пары денотатов. Таким образом, полная система классов семантического словаря образована множеством лексем, дополненных классами классификатора и классами бинарных базисных функций.

2.2. Словарь [4–6] имеет две формы представления. В исходной форме описание лексемы представлено в виде единственной формулы, объединяющей синтаксическую и семантическую информацию. В производной форме эта информация разнесена в два раздела. В синтаксическом разделе определяются условия

взаимодействия описываемой лексемы с другими лексемами предложения, ее синтактика. В семантическом разделе приводится формула толкования лексемы. Разделы жестко взаимосвязаны: синтаксический раздел содержит описание параметров формулы толкования. Информация в производной форме более удобна для использования в программах анализа текста, и далее именно ее мы будем считать основной.

2.3. Нас будет интересовать интерпретация толкований при выявлении кореферентных отношений в тексте. Для денотатов можно ввести уникальные системные обозначения, и считать эти обозначения "представителями" сущностей в системе – референтами референтных отношений. Обычно в информационных системах в качестве такого обозначения используется адрес, относительно которого в памяти системы собираются сведения, относящиеся к заданной этим адресом сущности. Тогда совокупность системных обозначений образует полное множество известных системе сущностей-референтов. Отметим, что в рамках концептуальных моделей данных классы рассматриваются как самостоятельные объекты, также представленные уникальными системными обозначениями.

2.4. Толкования лексем интерпретируются как отношения, определенные на системе классов семантического словаря [6]. Уже конструкция L$C (см. п.**2.1**.) фактически представляет собой запись отношения принадлежности класса L лексемы классу C из классификатора: любой денотат из класса лексемы принадлежит вместе с тем и заданному в имени лексемы классу. Записью *переменная* **:** *лексема* определяется область возможных (вне зависимости от контекста) денотативных значений указанной переменной: переменная может принимать значения из класса лексемы. В толкованиях используются следующие обозначения переменных:

переменная ::= тег_переменной номер_переменной

тег_переменной ::= **S||A||E||Y||V||Z||U||#**

номер_ переменной – одно- или двухразрядное число

Теги переменных дополняют описываемую другими средствами область возможных значений переменных морфосинтаксическими характеристиками значений:

S – существительное;

A – объект прилагательного;

E – объект наречия;

Y – субъект предлога;

V – глагол;

Z – семантический (валентный) актант лексемы, заданный ее синтактикой;

U – операнд союза;

– отмечает факт существования некоторого значения.

Примеры: A1; Z11; S0.

Толкование оформляется в виде суперпозиции отношений одного из следующих видов:

– операторная_переменная**:** лексема

– референтная_переменная **>**операторная_переменная**:**
лексема**(**список_актантов**)**

– референтная_переменная **>**операторная_переменная**:**
бинарная_функция**(**актант, актант**)**.

Операторная переменная определена на классе соответствующей лексемы или бинарной функции. Референтной переменной элемента указывается операторная переменная одного из актантов. Отношение в любом случае определяет область значений пары переменных: операторной и референтной. В записи отношения одна из переменных может явно не указываться.

2.5. Отношение является константным, если для него заданы значения переменных. Предикатом отношения в заданном контексте мы назовем функцию над отношением, принимающую значение "истина" лишь на имеющихся в этом контексте константных значениях отношения.

Толкование лексемы мы можем представить в виде конъюнкции предикатов терминальных отношений, полученных из отношения толкования по следующим правилам:

а) Конструкция R>O:*лексема***(***список актантов***)** преобразуется в конъюнкцию:

[R, O]:: O:*лексема*; O(*attr*$_1$**:** R$_1$); …; O(*attr*$_k$: R$_k$).

Здесь лексема принадлежит объектному классу, а

R – референтная переменная конструкции,

O – операторная переменная лексемы,

attr$_i$, R$_i$ – атрибут и референтная переменная i-того актанта.

Полученной конъюнкции в квадратных скобках сопоставлены результирующие операторная и референтная переменные.

Если на месте актанта задано отношение rel$_i$, где в качестве референтной используется операторная переменная O, то для него атрибутивное терминальное отношение не строится, а R>O:*лексема*(O:rel$_i$) в конъюнкции заменяется на O:rel$_i$.

Если переменная O в рассматриваемой конструкции не задана, список переменных толкования пополняется новым элементом – операторной переменной данной конструкции.

б) Та же конструкция R>O:*лексема***(***список актантов***)**, но лексема принадлежит к одному из классов-свойств, и в списке актантов указан объект-владелец свойства. Это может быть актант с атрибутом ОБЪЕКТ или актант, у которого в качестве референтной указана А-переменная:

$attr_1$:R_1> rel_1, … $attr_i$: A_i > rel_i, $attr_k$: R_k > rel_k).

В этом случае конструкция преобразуется в

[R>A_i:]:: A_i (*лексема*.O; O(*attr*1**:**R_1); …;O($attr_k$: R_k).

Отметим, что в результате преобразования меняется операторная переменная конструкции.

в) Конструкция O>S: *бинарная*(*актант*,*актант*) преобразуется в

[O,S]:: S: *бинарная*; S(R_1, R_2).

г) Преобразование очередной конструкции из суперпозиции толкования производится после того, как выполнены преобразования всех ее актантов. Результаты преобразования подсоединяются к ранее построенной конъюнкции.

2.6. Можно выделить четыре вида толкований в словаре:

А) К "синонимическим" толкованиям относятся те, у которых область определения операторной переменной тождественна классу денотативных значений толкуемой лексемы.

АВТОМАШИНА$1213241
Syn(S1:АВТОМОБИЛЬ$1213241(РОД:Z1, ДЛЯ:Z2))
Предикат толкования: Syn [S1]:: S1:АВТОМОБИЛЬ$1213241; S1((РОД:Z1); S1(ДЛЯ:Z2);

Б) Толкования, указывающие на принадлежность лексемы классу, не обозначенному в классификаторе явно. Как правило, этот вид толкований используется для существительных из класса физических объектов.

АВТОПОГРУЗЧИК$1213222
S1(* ОБЪЕКТ:Z1 *) S1:Usor(S1:МАШИНА$1213222,ПОГРУЗКА$152515(ОБЪЕКТ:Z1))
Предикат толкования:
дескрипция: [S1] :: S1: АВТОПОГРУЗЧИК$1213222; S1(* ОБЪЕКТ:Z1 *);
класс дескрипции: [S1] :: S1:МАШИНА$1213222; S2: ПОГРУЗКА$152515; S2(ОБЪЕКТ:Z1);S1(Usor:S2);

Интерпретация толкования: лексема АВТОПОГРУЗЧИК$1213222 принадлежит классу машин, предназначенных для погрузки. Предикат толкования имеет две части:

– дескрипцию, отношениями которой описываются класс денотативных значений лексемы и их возможные актанты;

– класс, подклассом которого является определяемая лексема.

В) Толкования лексем, значениями которых являются свойства, отличительные признаки тех или иных объектов. Этот вид толкований используется для объектов из классов свойств, для прилагательных, наречий и предлогов.

БЕЛЫЙ$12/0121 (классификация объекта)
Толкование: A1>S1:ЦВЕТ$12/012(A1,S1: БЕЛЫЙ $12/0121))
Предикат толкования:[A1, S1]:: S1: БЕЛЫЙ $12/0121; S1:ЦВЕТ$12/01; A1(ЦВЕТ:S1);

БЕЛЫЙ $12/0121 (классификация цвета или света)
Толкование: S1:ЦВЕТ$12/012(S1: БЕЛЫЙ $12/0121)
Предикат толкования :[S1]:: S1: БЕЛЫЙ $12/0121; S1:ЦВЕТ$12/01;

В этом виде толкований в качестве референтной переменной указываются либо A-, либо E-, либо Y-переменная. Операторная переменная не совпадает с референтной.

Г) Толкования, обеспечивающие референцию к денотативным значениям свойств объектов. Здесь в качестве операторной переменной указывается объект-владелец свойства, а в качестве референтной – описываемая лексема. В толкованиях этого вида владелец свойства – валентный актант определяемой лексемы и задан значением Z-переменной.

БЕЛИЗНА$12/1121 (объекта Z1)
Толкование: Z1>S1:ЦВЕТ$12/112(ОБЪЕКТ:Z1, S1:БЕЛЫЙ$12/1121)
Предикат толкования: [Z1] :: S1: БЕЛЫЙ $12/0121; S1:ЦВЕТ$12/01; Z1(ЦВЕТ:S1);

Отличительные признаки толкования:

– лексема относится к классу свойств;

– операторная переменная – валентный актант.

Д) Интерпретация толкований глаголов требует специального рассмотрения, которое вывело бы нас далеко за рамки данной статьи. Отметим лишь, что в [7] было показана целесообразность толкований глаголов дескрипциями с независимым денотативным статусом.

2.7. Рассмотрим конъюнкцию отношений R>O: *лексема*; O(*attr*: R_1) . Утверждение для каждого из денотативных значений O из класса *лексема* устанавливает свойственное ему значение R_1 атрибута *attr*:

R>O: *лексема*; O(*attr*: R_1) $\leftrightarrow$ $\forall$(O)($\exists$(R_1) (O: *лексема*; O(*attr*) $\rightarrow$ O(*attr*: R_1))

Имея терминальное отношение O: *лексема*; O(*attr*:R_1) и задавая конкретное значение O, мы имеем факт существования свойственных ему значений R_1. Иными словами, в нашем случае R_1 находится в сколемовой зависимости от O.

Если операторная переменная толкования является одновременно и референтной (варианты А, Б и Д), лексема имеет независимый денотативный статус: все прочие переменные толкования находятся в сколемовой зависимости от его операторной переменной. Вариант В толкований используется для лексем, имеющих подчиненный (квалификационный) статус: значения всех переменных толкования, в том числе и операторной, находятся в сколемовой зависимости от референтной переменной. Значения последней определяются той лексемой, к которой данная подсоединена в предложении подчиненной синтаксической связью (вид связи характеризуется тегом референтной переменной). Вариант Г – особый. Здесь операторная переменная – валентный актант лексемы, и именно она определяет значения прочих (в том числе и референтной) переменных толкования.

3. Предикаты предложений.

3.1. Мы будем исходить из следующих предположений относительно правильно построенных повествовательных предложений текста:

1. Предложению может быть сопоставлено дерево направленных от корня к листьям подчинительных связей между его словами, и для каждого из слов определена обозначенная этим словом лексема. В тройке

лексема 1 —связь—> лексема 2

параметром *связь* задается тождественность выходной переменной из толкования лексемы 1 и референтной переменной лексемы 2. В качестве выходной переменной лексемы 1 может быть указана либо валентная Z-переменная из толкования этой лексемы (валентная связь), либо ее операторная переменная (синтаксическая связь).

2. Заданное предложением дерево связей может быть разбито на поддеревья, каждое из которых представляется дескрипцией, имеющей независимый статус. Ее референтной переменной определяется обозначенное дескрипцией денотативное значение.

3. Порядок расположения дескрипций в общей конъюнкции предиката предложения не произволен: если референтная переменная дескрипции D1 входит в одно из отношений дескрипции D2, то D1 предшествует D2. Фактически, дескрипции вычисляются в порядке, заданном графом подчинительных связей в семантико-синтаксической модели предложения.

4. Дескрипция считается вычислимой, если в ее контексте для ее операторной переменной удается установить единственное денотативное значение. Предложение увязывается с ранее введенным текстом и информацией в базе знаний, если все его дескрипции вычислимы. Текст связен, если все его предложения увязаны. Таким образом, условия вычислимости каждой из дескрипций текста являются условиями связности этого текста.

Пример.

белая скатерть → [X1] :: X1:СКАТЕРТЬ$121386; S1:БЕЛЫЙ$12/1121; S1:ЦВЕТ$12/01; X1(ЦВЕТ:S1).

Здесь X1 – операторная (выходная) переменная дескрипции; она же является референтной.

Предикаты толкований лексем:

СКАТЕРТЬ

X1:СКАТЕРТЬ$121386 – отношение, задающее область определения операторной переменной.

БЕЛЫЙ

[A1, S1]:: S1: БЕЛЫЙ $12/0121; S1:ЦВЕТ$12/01; A1(ЦВЕТ:S1);

Связь лексем в предложении – синтаксическая, согласование прилагательного с существительным.

СКАТЕРТЬ$121386 —!Какой—> БЕЛЫЙ$12/1121

Переменная (X1) унифицируется с референтной входной (A1).

белизна скатерти → [S1, Z1] Z1:СКАТЕРТЬ$121386; S1: ЦВЕТ$12/112(ОБЪЕКТ:Z1, S1: БЕЛЫЙ$12/1121)

Толкование:

БЕЛИЗНА

[S1, Z1]:: S1: БЕЛЫЙ $12/0121; S1:ЦВЕТ$12/01; Z1(ЦВЕТ:S1);

Связь лексем в предложении – валентная:

БЕЛИЗНА$1/—Z1—>СКАТЕРТЬ$121386

3.2. В дескрипции значение ее операторной переменной O представлено конъюнкцией терминальных предикатов вида:

[R, O]:: O: *лексема*; O(*attr*$_1$: R_1); …;O(*attr*$_k$: R_k),

причем отношения атрибуции значений переменной O могут и отсутствовать. Общий запрос к контексту дескрипции относительно ее операторной переменной имеет вид:

"Найти значения переменной O в отношениях O: *лексема*"

Результатом запроса R={~o_1, ~o_2, …, ~o_n} являются все имеющиеся в контексте сущности (классы и экземпляры), удовлетворяющие условию запроса. Теперь из полученного списка значений необходимо вы-

брать одно, удовлетворяющее условиям дескрипции. Уточняющие запросы к контексту формируются из списка атрибутивных отношений дескрипции и имеют вид:

"Для сущности o_t ($1<o_t<n$) найти значения атрибутов R_i, …, R_j ($1\leq i, j\leq k$) в отношении О: *лексема*; О(*attr*$_i$ R_i); …; О(*attr*$_j$: R_j)".

Результат этого запроса – имеющиеся в контексте сущности из класса *лексема*, для которых имеются значения атрибутов R_i, … , R_j.

Возможны разные варианты отношения предиката О: *лексема*; О(*attr*$_1$**:**R_1); …; О(*attr*$_k$: R_k) к контексту дескрипции:

1. Относительно переменной О к началу ее вычисления известно, что ее значение отсутствует в контексте дескрипции. (Это может быть установлено, например, в результате анализа актуального членения предложения, его синтаксической структуры, морфологических свойств лексем предложения.) В этом случае контекст дескрипции необходимо пополнить новой сущностью – денотативным значением О. Значения переменных в атрибутивных отношениях переменной О находятся в сколемовой зависимости от нее и, следовательно, тоже являются новыми.

Пример: Кто-то в белом вошел в комнату.

В этом предложении неопределенным местоимением *кто-то* обозначен субъект действия, не известный автору текста, и, следовательно, не известный (по крайней мере, в момент анализа предложения) и интерпретатору текста тоже. Денотативное значение субъекта действия – новая сущность.

2. Относительно переменной О заведомо известно, что она обозначает сущность, ранее упоминавшуюся в тексте. При вычислении переменной О формируется общий запрос к контексту дескрипции на поиск этой сущности, и если ее найти не удается, то текст несвязен. При единственном ответе на общий запрос переменная О считается вычисленной. Если находится несколько возможных значений переменной, то формируются уточняющие запросы. Текст несвязен, если в результате подачи этих запросов выявить единственное значение переменной О не удается. Вычисленное значение О однозначно определяет значения переменных в ее атрибутивных отношениях, и эти значения могут оказаться как ранее заведенными в контексте дескрипции, так и новыми, вводимыми данной дескрипцией.

Пример: Автор (этой) книги был неизвестен читателям.

Здесь лексема КНИГА использована в роли описателя значения лексемы АВТОР, а словосочетание *автор книги* относится к теме предложения, что в совокупности позволяет с большой долей уверенности считать, что денотативное значение лексемы КНИГА – уже существующая в контексте данной дескрипции сущность. Местоимение *этой* вообще делает ситуацию однозначно трактуемой.

Предположим, что значение переменной О получено в результате выполнения уточняющего запроса. Тогда атрибутивные отношения дескрипции, не вошедшие в этот запрос и не означенные при вычислении предыдущих дескрипций предложения, считаются новой информацией и подлежат регистрации в контексте.

3. Вопрос о том, новая или уже существующая в контексте дескрипции сущность обозначена переменной О, может решаться только на основе запроса к этому контексту. При отсутствии ответа на общий запрос для переменной О в контексте дескрипции заводится новая сущность. Единственный ответ на общий и серию уточняющих запросов позволяет считать полученную сущность искомым значением О. В противном случае текст несвязен.

Пример: Близкий топот лошади заставил ее остановиться.

Относительно лексемы ЛОШАДЬ мы не можем с уверенностью утверждать, ни что ее значение является новым, ни что соответствующая сущность уже имеется в контексте.

3.3. Итак, нами определена конструкция дескрипции, обладающая следующими свойствами:

– дескрипция представляет собой конъюнкцию терминальных предикатов конкретизации и атрибуции;

– в списке переменных дескрипции выделена операторная переменная, подлежащая вычислению; прочие ее переменные находятся в сколемовой зависимости от выделенной;

– референтная переменная, значением которой является сущность, обозначенная дескрипцией, может отличаться от операторной, но в этом случае находится от нее в сколемовой зависимости;

– переменные дескрипции принадлежат общему списку предиката предложения;

– операторная переменная дескрипции не используется ни в одном из предикатов предшествующих дескрипций;

– значение операторной переменной дескрипции может зависеть от значений предыдущих дескрипций и не зависит от последующих.

4. Заключение

Цель нашей работы – из любого правильно построенного предложения русского языка извлечь утверждение в виде конъюнкции дескрипций, не накладывая при этом слишком жестких ограничений на условия правильности. Это решение определяется как денотативным статусом лексемы, т. е. ее способностью исполнять роль самостоятельной дескрипции в предикате утверждения, так и синтаксическим взаимодействием слова, представляющего лексему, с другими словами предложения.

Литература

1. Арутюнова Н.Д. Предложение и его смысл: логико-семантические проблемы. М.: Наука, 1976.

2. Степанов Ю.С. Имена. Предикаты. Предложения. М.: Наука, 1981.

3. Падучева Е.В. Высказывание и его соотнесенность с действительностью. М.: Наука, 1985.

4. Тузов В.А. Компьютерная семантика русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2004.

5. Тузов В.А. Компьютерная грамматика русского языка // Вестник С.-Петерб. ун-та. Сер.10. Прикладная математика, информатика, процессы управления. – СПб.: Изд-во С.-Петерб. ун-та, 2004. Вып. 1-2. – С. 94–100.

6. Лезин Г.В., Тузов В.А. Семантический анализ текста на русском языке: семантико-синтаксическая модель предложения // Экономико-математические исследования: математические модели и информационные технологии. – СПб: Наука, 2003. Вып. III. С. 282–303.

7. С.М. Герасимов, Г.В. Лезин. Компьютерный анализ текстов: извлечение формальных утверждений из повествовательных текстов. //Экономико-математические исследования: математические модели и информационные технологии. СПб: Наука, 2005. Вып. IV, Ч. II. С. 58–79.