УДК 519.7

Н.А. Валенда, Г.Ф. Дюбко

МЕТОДЫ АНАЛИЗА ЕСТЕСТВЕННОГО ЯЗЫКА НА ОСНОВЕ ФУНКЦИОНАЛЬНОЙ МОДЕЛИ СЕМАНТИКИ

Введение

Обработка информации на естественном языке является одной из фундаментальных задач в области искусственного интеллекта. В настоящее время интерес к этому направлению вызван необходимостью обработки постоянно увеличивающихся объемов информации на естественном языке. Это обусловлено распространением электронного документооборота, развитием информационных сетей. Обработка информации на естественном языке производится в системах автоматического перевода, поисковых системах, интерфейсах к базам данных, экспертных системах. Результаты работы систем обработки информации характеризуются достаточно низким качеством из-за сложности естественного языка как объекта моделирования. Повышение качества систем обработки информации на естественном языке является одной из актуальных задач в области искусственного интеллекта [1].

Одним из основных подходов к улучшению качества обработки информации является привлечение семантики и знаний о предметной области. Большинство современных работ направлены на построение формальных моделей семантики, которые позволят улучшить качество методов обработки информации. Целью данной работы является усовершенствование методов синтаксического и семантического анализа на основе модели семантики в виде семантических функций для повышения качества формального анализа языковых конструкций в системах искусственного интеллекта.

1. Анализ на основе атрибутных транслирующих грамматик

Следуя идеям Хомского, возможна формализация семантики языка – правил вычисления смысла предложения, с помощью его синтаксиса – правил построения предложений, которые задаются продукциями порождающей грамматики этого языка [2]. Для анализа формальных языков используют контекстно-свободные грамматики (КС). Естественные языки имеют значительно более сложную структуру, их анализ возможен только с учетом контекста. Для этого КС-грамматика расширяется за счет введения атрибутов.

Атрибутная грамматика — это расширение КС-грамматики, где каждому символу может быть приписано конечное число семантических параметров – атрибутов, которые принимают некоторые значения. Для каждого синтаксического правила определяются семантические правила, т.е. функциональные зависимости одних семантических атрибутов, приписанных символам, входящим в эту продукцию, от других атрибутов символов, входящих в ту же самую продукцию. Непосредственное вычисление значений всех атрибутов для каждого предложения языка определяется по дереву вывода этого предложения.

Дерево вывода в КС-грамматике – это помеченное упорядоченное дерево, каждая вершина которого помечена символом из множества $V_T \cup V_N \cup \varepsilon$. Если внутренняя вершина помечена символом S, а ее прямые потомки – символами $X_1,...,X_n$, то $S \to X_1,...,X_n$ – правило этой грамматики.

Атрибутной транслирующей грамматикой называется

$$G = (V_T, V_N, V_S, P, S),$$

где V_T – алфавит терминальных символов; V_N – алфавит нетерминальных символов; V_S – множество семантических правил; S – аксиома; P – множество продукций вида $A \to \alpha$, где

$$\alpha \in (V_T \cup V_N \cup V_S)^*.$$

В данной работе используются два типа семантических правил. Первый тип правил – это присвоение и проверка морфологических атрибутов. Для терминальных и нетерминальных символов правой части правил грамматики производится проверка грамматической сочетаемости на основании сравнения морфологических атрибутов. Если согласование выполнено, то производится присвоение морфологических атрибутов нетерминальному символу левой части правила, что соответствует формированию родительского узла в дереве вывода. Второй тип правил – это применение семантических функций для проверки смысловой сочетаемости дочерних вершин данного узла [3]. Эти правила оформляются в виде функций, которые включают механизмы семантического анализа. Слова словосочетания рассматриваются как аргументы семантической функции. Осмысленность языковой конструкции задается существованием семантической функции на заданных аргументах, т.е. значения аргументов должны принадлежать области определения семантической функции.

В случае выполнения семантического согласования, для данного фрагмента текста формируется семантическая функция, которая задает его смысл. Далее данный фрагмент рассматривается как единое целое и наследует морфологические атрибуты

главного слова. При выполнении правил обоих типов формируется новый узел, обладающий морфологической информацией и семантическим значением.

В общем виде правило грамматики имеет вид:

$$A_{x_1,\dots,x_n} \to B_{y_1,\dots,y_m} C_{z_1,\dots,z_k} F_i(y_1, z_1),$$

$$x_1 = y_1, \dots, x_n = y_m.$$

где A, B, C – нетерминальные символы; $x_1,\dots,x_n$, $y_1,\dots,y_m$, $z_1,\dots,z_k$ – морфологические атрибуты; $F_i(y_1, z_1)$ – функция семантического согласования.

Содержательно символы из различных алфавитов имеют следующий смысл: символы терминального алфавита являются словами анализируемой языковой конструкции; нетерминальные символы – грамматические категории языка; семантические правила – проверка грамматической и семантической сочетаемости элементов языковой конструкции.

Терминальный символ имеет следующий вид:

$$X_{a_1,\dots,a_n},$$

где X– имя терминального символа; $a_1,\dots,a_n$ – атрибуты терминального символа.

В качестве имени терминального символа рассматривается часть речи: существительное – *С*; прилагательное – *П*; причастие – *Пр*; числительное – *Чс*.

Атрибутами являются морфологические признаки, a_1 всегда задает слово. Остальные атрибуты соответствуют признакам из комплекса морфологической информации. Для существительных: одушевленность, род, число, падеж. Для прилагательных: род, число, падеж, степень сравнения, краткость. Для причастий: род, число, падеж, краткость, вид, время, залог. Для числительных: род, число, падеж, тип.

Традиционно синтаксический анализ производит разбор текста по предложениям. Строится дерево разбора предложения на основе вывода в грамматике, которое передается на вход семантическому анализатору для дальнейшего анализа. В предложениях русского языка порядок слов не является строгим, что значительно усложняет задачу синтаксического анализа. Введение дополнительных правил делает грамматику громоздкой и малоэффективной. Если рассмотреть более мелкие единицы текста – словосочетания, то можно сделать вывод, что порядок слов в них является более строгим. Структура словосочетаний значительно проще структуры предложений. Для упрощения грамматики и последующего анализа предлагается сузить область применения синтаксического анализа с предложения до словосочетания. Это позволит реализовать анализ без возвратов.

Метод анализа на основе атрибутных транслирующих грамматик основывается на реализации вывода языковых конструкций в атрибутной грамматике. Анализатор строится на основе LR(1)–грамматики. На вход анализатора подаются простые предложения естественного языка. На выходе получаем предложение, разбитое на слова и словосочетания. В качестве анализируемых конструкций выступают словосочетания. Анализ осуществляется в несколько проходов. Каждый проход линейно зависит от объема входных данных, поэтому общая функция зависимости времени работы от объема входных данных имеет линейный характер.

В работе рассматриваются словосочетания, имеющие в качестве главного слова глагол или существительное. На примере словосочетаний с главным словом существительным (именных групп) рассмотрим работу метода анализа. Именная группа состоит из существительного и зависимых слов, в качестве которых могут выступать как существительные, так и прилагательные, причастия, порядковые числительные. Аргументами именной группы могут быть как отдельные слова, так и слова уже объединенные в именную группу. Из этого следует, что правила, допускающие именные группы, должны быть рекурсивными.

К рекурсивным правилам формирования именных групп относятся:

— сочетание существительного с прилагательным, причастием, числительным;

— сочетание существительного с существительным в родительном падеже.

К первой категории относятся словосочетания такого вида – «новые тонированные окна», ко второй – «портфель профессора информатики». Несмотря на различную структуру обоих словосочетаний мы имеем дело с правой рекурсией. Чтобы убедиться в этом, представим эти словосочетания в виде скобочных записей:

(новые (тонированные окна);

(портфель (профессора информатики).

Для образования именных групп важен порядок применения правил. Сначала необходимо создать именные группы с прилагательными, причастиями и числительными. В дальнейшем они могут входить в состав именных групп родительного падежа. Поэтому, сначала строятся нетерминалы – <*Именная группа_1*>, а затем <*Именная группа_2*>. Нетерминалы <*Именная группа_1*> рассматриваются как терминальные символы при втором проходе анализатора. Словосочетания заменяются главными словами, чтобы упростить анализ. Вся информация сохраняется и восстанавливается после допуска языковой конструкции.

Грамматика, распознающая словосочетания с главным словом существительным, приведена ниже.

1.1 $<$*Именная группа_1*$>_{r_1,\dots,r_5} \to$

$\Pi_{p_1,\dots,p_6} < A >_{p'_1,\dots,p'_5} F_1(p_1, p'_1)$

$r_1 \leftarrow p'_1, \dots, r_5 \leftarrow p'_5$

1.2 $<$*Именная группа_1*$>_{r_1,\dots,r_5} \to$

$Пр_{p_1,\ldots,p8} < A >_{p'_1,\ldots,p'_5} F_2(p_1, p'_1)$

$r_1 \leftarrow p'_1, \ldots, r_5 \leftarrow p'_5$

1.3 $<$*Именная группа_1*$>_{r_1,\ldots,r_5} \rightarrow$

$Чс_{p_1,\ldots,p_5} < A >_{p'_1,\ldots,p'_5} F_3(p_1, p'_1)$

$r_1 \leftarrow p'_1, \ldots, r_5 \leftarrow p'_5$

2.1 $< A >_{r_1,\ldots,r_5} \rightarrow C_{p_1,\ldots,p_5}$

$r_1 \leftarrow p_1, \ldots, r_5 \leftarrow p_5$

2.2 $< A >_{r_1,\ldots,r_5} \rightarrow <$*Именная группа_1*$>_{p'_1,\ldots,p'_5}$

$r_1 \leftarrow p'_1, \ldots, r_5 \leftarrow p'_5$

3 $<$*Именная группа_2*$>_{r_1,\ldots,r_5} \rightarrow C_{p_1,\ldots,p_5}$

$<$*Именная группа_p*$>_{p'_1,\ldots,p'_5=rod} F_4(p_1, p'_1)$

$r_1 \leftarrow p_1, \ldots, r_5 \leftarrow p_5$

4.1 $<$*Именная группа_p*$>_{r_1,\ldots,r_5} \rightarrow \varepsilon$

$r_1 = 0, \ldots, r_5 = 0$

4.2 $<$*Именная группа_p*$>_{r_1,\ldots,r_5} \rightarrow C_{p_1,\ldots,p_5=rod}$

$<$*Именная группа_p*$>_{p'_1,\ldots,p'_5} F_5(p_1, p'_1)$

$r_1 \leftarrow p_1, \ldots, r_5 \leftarrow p_5$

Приведем пример вывода в данной грамматике языковой конструкции: «портфель профессора информатики». Для каждого слова в результате морфологического анализа получим множество терминалов V_T:

портфель – $C_{\text{нод, муж, ед, им}}$, $C_{\text{нод, муж, ед, вин}}$;

профессора – $C_{\text{од, муж, ед, род}}$, $C_{\text{од, муж, ед, вин}}$, $C_{\text{од, муж, мн, им}}$;

информатики – $C_{\text{нод, жен, ед, род}}$.

В табл. 1 приведен вывод допустимой строки с использованием стека.

В результате вывода получатся две допустимые последовательности терминальных символов за счет того, что слово «портфель» может быть как в именительном, так и в родительном падеже. Эту неоднозначность невозможно снять на данном этапе даже на уровне семантики.

Вывод, приведенный выше, не рассматривает вычисление семантических функций, которые являются семантическими правилами в грамматике. Рассмотрены только правила для согласования грамматических атрибутов. Выполнение семантического согласования будет приведено в семантическом анализе.

Таблица 1

Вход	Стек	Атрибуты
$C_{\text{нод., муж, ед, им/вин}}$ $C_{\text{од., муж, ед, род}}$ $C_{\text{нод., жен, ед, род}}$		
$C_{\text{од., муж, ед, род}}$ $C_{\text{нод., жен, ед, род}}$	C_1	
$C_{\text{нод., жен, ед, род}}$	C_1C_2	1. нод, муж, ед, им/вин. 2. од, муж, ед, род.
	$C_1C_2C_3$	1. нод, муж, ед, им/вин. 2. од, муж, ед, род. 3. нод, жен, ед, род.
	$C_1C_2C_3$ $<ИГ_p>_4$	1. нод, муж, ед, им/вин. 2. од, муж, ед, род. 3. нод, жен, ед, род. 4. 0, 0, 0, 0.
	$C_1C_2 <ИГ_p>_3$	1. нод, муж, ед, им/вин. 2. од, муж, ед, род. 3. нод, жен, ед, род.
	$C_1 <ИГ_p>_2$	1. нод, муж, ед, им/вин. 2. од, муж, ед, род.
	$C_1 <ИГ_p>_2$	1. нод, муж, ед, им/вин. 2. од, муж, ед, род.
	$<ИГ_2>_1$	1. нод, муж, ед, им/вин.
	допуск	

2. Семантический анализ

Использование функционально семантического подхода основано на применении семантического словаря, содержащего значения слов, которые приписываются им в языке. Такую информацию можно получить из толкового словаря. Для формализации значений слов используется аппарат семантических функций [4]. В электронном словаре каждому слову соответствует отдельная запись, представляющая собой множество суперпозиций семантических функций. Каждая суперпозиция соответствует одному значению слова.

На этапе семантического анализа каждое слово должно получить однозначную интерпретацию на всех уровнях — морфологическом, синтаксическом, семантическом. Для каждого слова должно быть указано, в какой роли оно выступает в предложении – объекта, субъекта, действия и т.д., а для многозначных слов – в каком из своих значений используется в предложении.

На вход семантического анализа конструкции естественного языка подаются в виде фрагментов, которые были получены на этапе синтаксического анализа. Каждый фрагмент состоит из набора лексем. Лексемами могут являться как слова, так и словосочетания. На этапе синтаксического анализа все слова языковой конструкции должны быть сгруппированы вокруг существительных и предикативных частей речи, кроме того, могут встречаться различные смысловые связки, которые не будут включаться ни в одну группу. Каждый набор лексем должен содержать одно предикативное слово. В противном случае, этот фрагмент должен быть разобран как словосочетание еще на этапе синтаксического анализа. Задача семантического анализа

состоит в заполнении пустых позиций предиката словами и словосочетаниями данного фрагмента. Словосочетания, образованные в результате синтаксического анализа, прошли семантическое согласование и для них созданы семантические функции.

Анализ конструкций естественного языка предполагает, что на основе языкового представления в результате некоторой последовательности действий будет построено внутреннее представление на формальном языке, которое будет отображать значение исходной языковой конструкции. Записи семантического словаря представляют собой формализованное описание значения, которое является шаблоном, позволяющим включать в формулу зависимые слова. Каждая функция вырабатывает значение, которое таблично сопоставлено ей в словаре. На основе значения функции ее можно встраивать в незаполненные позиции другой функции, получая суперпозицию функций. На основании этих свойств построен метод семантического анализа, позволяющий устанавливать зависимости между лексическими единицами в предложении. Этот метод является развитием метода семантического анализа на основе фрейма предиката. В данной работе фрейм предиката задается суперпозицией семантических функций, соответствующих предикатному слову в словаре.

Семантический анализ состоит из двух частей:

— проверка семантики на уровне синтаксического анализа;

— формирование функции для языковой конструкции на основе предиката.

Проверка связи, установленной на уровне синтаксического анализа, производится при формировании словосочетаний. Вызов процедуры семантического анализа задается функцией $F_i(x_1, x_2)$, где x_1 — главное слово словосочетания, а x_2 — зависимое, i — определяет на основе функции главного или зависимого слова будет ли строиться функция словосочетания, а также стандартную функцию.

В терминологии грамматик можно сказать, что это анализ снизу вверх — на основании простых единиц строятся более сложные, пока не будет получено предложение. Простейшими составляющими являются слова. Для объединения слов в словосочетания используются семантические функции, соответствующие им в словаре. Основой метода формирования простых словосочетаний является заполнение пустых позиций в суперпозициях семантических функций. На основании простых словосочетаний формируются сложные словосочетания. Основой для формирования словосочетаний является синтаксический анализ, что позволяет не делать полный перебор всех возможных сочетаний слов. Если в предложении есть предикативная часть речи, то полученные словосочетания и слова, которые не вошли в словосочетания, встраиваются в функцию предиката.

Метод анализа состоит из трех частей:

— формирование простых словосочетаний;

— формирование сложных словосочетаний;

— заполнение функции предиката.

Две первые задачи решаются на этапе синтаксического анализа, третья — на этапе семантического анализа.

Формирование простого словосочетания.

1. На основании синтаксического анализа определяется главное и зависимое слово словосочетания – x_1, x_2. Из семантического словаря извлекаем функции, соответствующие главному и зависимому словам $f_i^{x1}\left(x_i^1,...,x_j^k\right)=\alpha_1$, $f_j^{x2}\left(x_k^1,...,x_l^n\right)=\beta_1$. Функция словосочетания строится на основе функции зависимого слова для словосочетаний существительное-прилагательное и на основе функции главного слова для словосочетаний существительное-существительное. В первом случае функцией словосочетания будет являться f_j^{x2}, а во втором — f_i^{x1}. Для заполнения переменных будет использоваться значение противоположной функции.

2. Если аргумент функции словосочетания x_m^k является переменной – $X_{[P_{morf},P_{sem}]}$, подставляем вместо него значение γ равное α_1 или β_1. Для P_{sem} и γ определяется близость выражаемых понятий: они должны либо совпадать как словарные константы, либо в иерархии значений P_{sem} должно быть более общим понятием, чем γ. Если P_{sem} и γ удалось сопоставить, то осуществляется подстановка $X_{[P_{morf},P_{sem}]}/V(x_1)$ или $X_{[P_{morf},P_{sem}]}/V(x_2)$ и выдается функция словосочетания, которая является суперпозицией функций главного и зависимого слова.

3. Если никакой аргумент x_m^k функции словосочетания не является переменной, то формируется стандартная функция на основе синтаксической связи. Ее проверка осуществляется на основании унификации значений входящих в нее слов. Их значения должны пересекаться. Если пересечения нет, то словосочетание не создается.

Формирование сложного словосочетания.

1. В качестве составных частей выступают простые словосочетания и слова. Для анализа используются главные слова простых словосочетаний. На основании синтаксического анализа определяется главное и зависимое слово словосочетания по тому же принципу, что и для простых словосочетаний. Выделяется функция словосочетания f_j^{xi}. Заполнитель γ формируется на основании значения главного слова словосочетания или значения одинарного слова.

2. Если аргументом функция f_j^{xi} является переменная $X_{[P_{morf},P_{sem}]}$, подставляем вместо нее γ. Если P_{sem} и γ удалось сопоставить, то осуществляется подстановка. Если x_γ является главным словом словосочетания, то подстановка имеет вид $X_{[P_{morf},P_{sem}]}/f_j^{x\gamma}$, если x_γ — одинарное слово, то $X_{[P_{morf},P_{sem}]}/V(x_\gamma)$.

3. Если функция зависимого слова не содержит переменных, то формируется стандартная функция на основе синтаксической связи. Ее проверка осуществляется на основании унификации значений входящих в нее слов. Если пересечения нет, то словосочетание не создается.

Формирование языковой конструкции на основе заполнения позиций предиката.

1. Из словаря берется функция, соответствующая предикативному слову, создается список переменных $\{X_{1[P_{morf},P_{sem}]},\ldots,X_{n[P_{morf},P_{sem}]}\}$.

2. Среди главных слов, сформированных словосочетаний, и одинарных слов ищутся те, что удовлетворяют P_{morf}, P_{sem}, переменных X_i.

3. Если существует несколько вариантов для заполнения X_i, то на основании методов снятия многозначности выбирается один из вариантов.

4. Если все слова языковой конструкции не могут быть включены в одну формулу, являющуюся суперпозицией функций, то данный отрезок текста семантически не связан. Если для предиката существует несколько значений, то переходим к следующей формуле.

5. Если все переменные предиката

$$\{X_{1[P_{morf},P_{sem}]},\ldots,X_{n[P_{morf},P_{sem}]}\}$$

заполнены или использованы все заполнители, то возвращается текущая функция предиката – предложение разобрано.

На основе этих методов разработан алгоритм семантической проверки синтаксической связи. Результатом работы семантической проверки может быть:

— формирование функции словосочетания, объединение лексем исходных слов в лексему словосочетания;

— лексемы исходных слов остаются без изменений, словосочетание не формируется.

Рассмотрим семантическое согласование на уровне синтаксического анализа на примере языковой конструкции «портфель профессора информатики». При анализе словосочетания функция смыслового согласования вызывается дважды. В первый раз при формировании нетерминального символа $<ИГ_р>_3$ вызывается F_5 (*профессора, информатики*).

Из словаря извлекаются семантические функции.

V_1 (*профессор*)= f_3 (*профессор*, Y) = *преподаватель*;

Y – *учебное заведение / научное заведение / учебное подразделение / наука.*

V_2 (*профессор*) = f_3 (*профессор*, Y) = *ученое звание*;

Y – *фамилия.*

V_1 (*информатика*) = *наука.*

V_2 (*информатика*) = *дисциплина.*

Для переменной Y формируется множество заполнителей :

$$M_{Y_{V_1}} = \{V_1\,(информатика), V_2\,(информатика)\}.$$

Производится сопоставление семантических признаков переменной Y с множеством заполнителей.

P_{semY} = *учебное заведение / научное заведение / учебное подразделение / наука*;

V_1 (*информатика*) = *наука*;

$P_{semY} = V_1$ (*информатика*).

Подстановка:

Y / V_1 (*информатика*).

Функция словосочетания:

f_3 (V_1 (*профессор*), V_1 (*информатика*)).

При формировании нетерминального символа $<ИГ_р>_2$, который задает все словосочетание в целом, происходит аналогичный процесс. В результате, получаем семантическое представление всего словосочетания:

f_3 (V (*портфель*), f_3 (V_1 (*профессор*), V_1 (*информатика*))).

Слово портфель имеет единственное значение, поэтому для его функции не указывается индекс значения. Данное формальное представление является результатом работы семантического анализатора. В нем каждое слово представлено единственным значением, которое задается индексом функции V. Связи между словами задаются функциями f_3. Семантический анализ на основе предиката не требуется.

Выводы

В данной работе получили дальнейшее развитие методы анализа на основе атрибутных транслирующих грамматик, благодаря введению семантических функций в виде символов действия, что позволило проверять существование семантического значения для словосочетаний, установленных на уровне синтаксиса. Усовершенствован метод анализа на основе фрейма предиката, благодаря использованию логического вывода на семантических функциях, что позволяет сопоставлять многозначным словам единственное значение.

Список литературы: 1. *Широков В.А.* Феноменологія лексикографічних систем. К.: Наукова думка, 2004. — 327 с. 2. *Карпов Ю.Г.* Основы построения трансляторов. СПб.: БВХ-Петербург, 2005. — 272 с. 3. *Валенда Н.А., Дюбко Г.Ф.* Принципы реализации анализа естественного языка // Проблемы бионики. № 50. 1999. — С. 45–53. 4. *Валенда Н.А.* Модель функционально семантической обработки текстов на основе унификации // Восточно-Европейский журнал передовых технологий. № 4/2(16). 2005. — С. 95–99.

Поступила в редколлегию 23.11.2005