

А.В.Сокирко

РГГУ

sokirko@diatech.ru

Введение

В настоящей статье описывается алгоритм, строящий первичное семантическое представление предложений русского языка в прикладной системе Диалинг. Система Диалинг существует уже больше года на базе ООО Диатех. Диалинг – это система русско-английского машинного перевода. Ее создание началось не на пустом месте. Основные концепции и аппарат автоматического анализа русских текстов были разработаны Н.Н.Леонтьевой (см. [1] и др.) и частично реализованы в системе «Политекст» [2], разрабатываемой лингвистической группой Центра Информационных Исследований (ЦИИ, который базировался в 1991-98 гг. в Институте США и Канады РАН). В разное время к участию в работе над системой Политекст привлекались студенты и аспиранты РГГУ. Некоторые из них в настоящее время работают над системой Диалинг и продолжают развивать (в части русского анализа) те направления, которые были «заморожены» или отложены в составе ЦИИ.

Автору данной статьи принадлежала программная реализация компонента графематического анализа и создание последней работающей версии базы данных РОСС (см. ниже), эти компоненты до сих пор функционируют в НИВЦ МГУ. В системе Диалинг они были воссозданы на новой платформе, переориентированы на задачу машинного перевода, время работы программ на порядок ускорено. Морфологический анализ воссоздан с небольшими изменениями, синтаксический анализ был развит и запрограммирован, а компонент семантического анализа полностью реализован автором. Все это не только позволяет решать конкретные практические задачи (улучшение качества машинного перевода), но и создает базу для исследовательских проектов. Так, результаты семантической интерпретации позволяют проследивать некоторые свойства исходного текста, например, смысловой неполноты. (Сошлемся на поддерживаемые грантами Российского Фонда фундаментальных исследований исследовательские проекты РФФИ-97: 97-06-80093 и РФФИ-99: 99-06-80296, в которых автор участвовал в 1997-99 гг.).

С учетом сказанного вернемся к первичному семантическому анализу, который на основе результатов синтаксического анализа и с использованием нескольких словарных комплексов строит первичное семантическое представление. На вход алгоритма первичного семантического анализа подаются результаты синтаксического и фрагментационного анализов, выполненных коллективом

системы Диалинг, формат и содержание которых будут описаны ниже. На выходе мы имеем первичный семантический граф одного предложения. Описываемый алгоритм использует в своей работе следующие семантические словари:

- 1) Русский общесемантический словарь (РОСС);
- 2) Словарь устойчивых оборотов;
- 3) Словарь групп времени;
- 4) Словарь локативных слов;
- 5) Общетематический тезаурус;
- 7) Финансовый тезаурус.

Подробное описание первых двух словарей можно найти в работах [1, 3 и др.], где в основном задан и аппарат используемого нами семантического языка и представления. Графематический и морфологический анализы почти полностью совпадают с аналогичными анализами в системе ПОЛИТЕКСТ. Описание синтаксического и фрагментационного анализов будет дано в отдельной статье.

Первичный семантический анализ мы далее называем семантическим, поскольку это первый из этапов системы Диалинг, который использует валентную структуру, записанную в словаре РОСС. Семантический анализ строит семантическую структуру одного предложения на русском языке. Семантическая структура состоит из семантических узлов и семантических отношений. Это центральные понятия используемой семантической теории, которую мы взяли за основу (см. подробнее [1-2]). Интуитивное определение узла формулируется примерно так:

Семантический узел – это такой объект текстовой структуры, у которого заполнены все валентности, как эксплицитно выраженные в тексте, так и имплицитные – те, которые получаются из экстралингвистических источников.

Из определения следует, что окончательный семантический узел может быть построен только в самом конце семантического анализа. Собственно говоря, главная цель семантического анализа – построение семантических узлов, которое подразумевает заполнение всех валентностей.

Кроме семантических узлов, существуют их атрибуты. Проблема заключается в том, как отличить одно от другого. Семантические узлы строятся из слов исходного предложения. Главный источник гипотез о составе семантического узла дает, безусловно, синтаксический анализ. Многие синтаксические группы могут перейти в семантические узлы, другие должны превратиться в атрибуты узлов. Другими источниками являются словарь временных групп, РОСС и все тезаурусы.

Однако, все вышесказанное – теория, а практика (алгоритмические определения тех же объектов) лишь приближается к этой теории. Ниже мы опишем нашу реализацию части этой теории.

С точки зрения реализации, например, очень важно, образуют ли слова, вошедшие в сем. узел, непрерывный отрезок в предложении или нет. Все семантические узлы, которые получены из синтаксиса, тезаурусов и словаря групп времени, неразрывны. Но семантические узлы, образованные из значений лексических функций-параметров, могут быть разрывны. Например, предложение *Иванов отдал солдатам приказ* содержит три семантических узла:

- 1) *Иванов*
- 2) *отдать приказ*
- 3) *солдаты*

2. Вход семантического анализа

- 1) Синтаксическое представление одного предложения на русском языке,
- 2) Все семантические словари и тезаурусы, привлеченные к реализации в системе Диалинг.

3. Выход семантического анализа

ИДЕНТ (*Пашков, дом*)
ó ОБ (*дом, стоять*)
ЛОК (*на пригорке ,*
стоять)

Множество семантических структур, построенных на основе входного синтаксического представления. Например, для предложения *Дом Пашкова стоит на пригорке* будет построена следующая семантическая структура:

Множество семантических структур должно быть упорядочено по степени правдоподобности. Кроме слов, семантические узлы могут включать:

- 1) Лексические функции [6]. Например, для предложения *Снижение составило восемь процентов*:
- 3) Множественный актант (МНА). Например, для предложения *Он смотрел на меня и на Петра*:

3) Связку (Copul). Например, для предложения *Поле представлялось ему бесконечным*:

Алгоритм

I. Инициализация семантических узлов

Для каждого фрагмента берем лучшие синтаксические варианты по размеру покрытия (покрытие - число слов, вошедших в синтаксические группы).

Для каждой группы или одиночного слова создаем узел, в который переносим всю необходимую информацию (из каких слов состоит, какие граммы у группы, главное слово группы, тип группы). Эту информацию мы будем называть синтаксическими характеристиками семантического узла.

Следуя определениям, данным в описании синтаксического анализа, назовем группу максимальной, если она не входит ни в одну другую группу.

Для максимальных групп типа ГЕНИТ_ИГ (именная группа, управляющая другой ИГ в генитиве) и ПРЯМ_ДОП узел не создается, так как в их основе лежит валентное отношение, информация о заполнении которого необходима в семантическом анализе.

Максимальная группа типа ОДНОР_ИГ также не преобразуется в узел, поскольку семантический анализ использует более общие механизмы сборки однородных членов, впрочем, информация о существовании такой группы все-таки используется.

II. Получение семантических вариантов по номерам значений

Поскольку в РОССе слово может иметь несколько значений, а во фрагмент может входить много слов с неединственным значением, при анализе необходимо перебрать все варианты интерпретации слов номерами значений. Например, если фрагмент состоит из двух слов, у каждого слова - по два значения, то число семантических вариантов по номерам значений будет $2 \times 2 = 4$.

Каждый семантический вариант по номерам значений рассматривается отдельно. Потом из них выбираются лучшие (ср. с [5]).

Каждый вариант по номерам значений оценивается только после выполнения следующих процедур:

- 1) вычисление семантически главных слов в синтаксических группах;
- 2) построение узлов в кавычках;
 - 3) построение узлов по словарю групп времени;
 - 4) построение узлов типа “друг друга”;
 - 5) построение узлов “краткое прилагательное+связка”;
- 6) подключение операторов типа “не”, “только”;
- 7) построение семантических узлов по лексическим функциям типа Magn;

- 8) построение семантических узлов по лексическим функциям типа Oper;
- 9) объединение локативных узлов;
- 11) интерпретация тезаурусных должностей и организаций;
- 12) интерпретация двойных союзов внутри фрагмента;
- 13) процедура построения всех отношений на основе грамматических характеристик (ГХ);
- 14) построение узлов типа МНА;
- 15) процедура построения и вычисления веса вариантов деревьев.

Все варианты деревьев включаются в выходное множество семантических структур предложения, причем варианты упорядочиваются по весу, который вычисляется в последней процедуре.

Ниже будут описаны некоторые из вышеперечисленных процедур.

III. Вычисление семантически главных слов в синтаксических группах

Семантически главное слово узла – это представитель узла в структуре, так, в текущей реализации программы все валентности узла берутся от семантически главного слова.

Семантически главное слово узла равно:

- 1) синтаксически главному слову, если тип группы, на основе которой он образован, не является ПГ (предложной группой) или ЧИСЛ-СУЩ;
- 2) если тип группы, из которой образован узел, - ПГ, то семантически главным словом считается семантически главное слово узла, который является непосредственным потомком данного;
- 3) если тип группы, из которой образован узел, - ЧИСЛ-СУЩ, то семантически главным словом считается семантически главное слово узла, который следует за первой непосредственно входящей в группу подгруппу.

Например, главное слово узла 'красный помидор' – помидор, узла 'в лесу' – лес, 'два дома' - дома.

IV. Процедура построения всех отношений на основе грамматических характеристик

Поскольку в основной своей массе узлы создаются по синтаксическим группам, некоторые узлы могут полностью (по словам) входить в другие. Назовем максимальным узлом узел, который не входит ни в один другой узел.

Для каждого максимального узла проводим все связи, которые можно провести к другим максимальным узлам, основываясь на словарных статьях главных слов узлов и синтаксических характеристиках узлов. Учитываются только ГХ_i (грамматические характеристики актантов) и некоторые очевидные грамматические свойства семантических отношений. К последним относятся:

- 1) Обязательное заполнение зависимого узла отношения ИДЕНТ узлом, первое слово которого обладает графематической пометой Бб, ББ, ЦК или ЦБК (графематические пометы: слово, начинающееся с большой буквы, состоящее из одних больших букв, цифровой и цифро-буквенный комплекс);
- 2) Возможное заполнение валентности КОЛИЧ актантом с пометой ЦК или построенным на группе КОЛИЧ, ОДНОР_ЧИСЛ.

Для того, чтобы описать то, как сопоставляются ГХi с некоторым узлом нужно ввести некоторые обозначения:

1) ЧАСТЬ_РЕЧИ(У)=Х обозначает, что семантически главному слову узла У приписана часть речи Х (морфологическая информация).

2) $X \hat{I} \text{ГРАММЕНЫ}(У)$ обозначает, что семантически главному слову узла У приписаны граммема Х(морфологическая информация);

3) $X \hat{I} \text{ГР_ДЕСКР}(У)$ обозначает, что Х – один из графематических дескрипторов семантически главного слова У (графематическая дескрипция).

4) $X = \text{ТИП_ГРУППЫ}(У)$ обозначает, что узел У образован из синтаксической группы Х.

Ниже будет определен предикат $P(У1, ГХi, У2)$, который равен 1, если узел У2 удовлетворяет значению поля ГХi, входящего в словарную статью семантически главного слова узла У1 (в противном случае предикат равен 0).

Итак, $P(У1, ГХi, У2)$ равен:
если ГХi = подл:*

то если ЧАСТЬ_РЕЧИ(У1)=INFINITIVE

то 0;

иначе если ГХi = подл:инф

то ЧАСТЬ_РЕЧИ(У2)=INFINITIVE;

иначе им $\hat{I} \text{ГРАММЕНЫ}(У2)$;

иначе если (ГХi = r:*)

и ($r \hat{I} \{Д_СИН_О, Д_ГРУППЫ\}$)

то если (ГХi = r:c)

и ($c \hat{I} Д_ПАДЕЖИ$)

то $c \hat{I} \text{ГРАММЕНЫ}(У2)$

и (ЧАСТЬ_РЕЧИ(У2) $\hat{I} \{NOUN, NOUN_g, NOUN_n, PRONOUN, PRONOUN_P, ADJ_FULL\}$

или ИЛЕ $\hat{I} \text{ГР_ДЕСКР}(У2)$

)

иначе если (ГХi = r:инф)

то ЧАСТЬ_РЕЧИ(У2)=INFINITIVE;

иначе если (ГХi = r:p+c)

и ($r \hat{I} \{Д_СИН_О, Д_ГРУППЫ\}$)

и ($p \hat{I} Д_ПРЕДЛОГИ$)

и ($c \hat{I} Д_ПАДЕЖИ$)

то (У2 = x1+x2)

и ($x1 = p$)

и ($x2$ – узел)

и ($c \hat{I} \text{ГРАММЕНЫ}(У2)$)

иначе если (ГХi = r:ИЛЕ)

то ИЛЕ $\hat{I} \text{ГР_ДЕСКР}(У2)$

иначе =>

=> если (ГХi = прим_опр:ИГ)
то У1 и У2 согласованы по числу и падежу;
или ИЛЕ Î ГР_ДЕСКР(У2)

иначе если (ГХi = ОБСТ_ГР:НАР)
то ЧАСТЬ_РЕЧИ(У2) = ADV

иначе если (ГХi = ОБСТ_ГР)

то ТИП_ГРУППЫ(У2) = ПГ
или ЧАСТЬ_РЕЧИ(У2) = ADV

Процедура построения всех отношений на основе грамматических характеристик почти всегда строит циклический граф, т.е. отношений в нем гораздо больше, чем на самом деле должно быть. Например, если у узла У1 есть валентность на именную группу в родительном падеже, то проводятся все связи от этого узла к именным группам в родительном падеже.

Получается так, что эта процедура строит все возможные грамматические решения. Задача дальнейших этапов - оценить и упорядочить по перспективности построенные гипотезы. Фактически, построенный граф - это таблица возможного управления, как она была введена [5] в так называемом методе фильтров.

V. Интерпретация однородных именных групп

Для каждой однородной именной группы, собранной синтаксисом, создается узел МНА (множественный актант). Пусть Х – последний именной узел однородного ряда, который собрал синтаксис. Х всегда будет членом МНА, поскольку перед ним стоит союз (оператор МНА). Пусть

Все падежи, которыми может управлять первый
предлог У, если У - предложная группа

Все падежи, которые приписаны У, если У – беспредложная ИГ

$F(Y) =$

{

Пусть $M = \{Y \mid (F(X) \subset F(Y) \cap \mathcal{A}) \ \& \ \exists A ((F(A) \subset F(Y) \cap \mathcal{A}) \ \& \ (A \neq X) \ \& \ (F(X) \subset F(Y) \cap \mathcal{A}))\}$. M – это множество потенциальных членов однородного ряда. В простейшем случае множество M состоит из согласованных с узлом X по падежу узлов.

Дальше для каждого отношения R , идущего из некоторого узла Z в какой-нибудь элемент M , создаем отношение из Z в только что созданный узел типа МНА. Характеристики нового отношения равны характеристикам отношения R

Таким образом, создаются гипотезы о том, какие отношения входят в узел МНА и что может подчинять узел МНА

VI. Процедура построения вариантов деревьев[1]

Пусть G – построенный на предыдущем этапе граф. Пусть $T_1, \dots, T_n$ – все остовные деревья графа G [2], в котором нет двух семантических отношений с одинаковым непустым названием, идущих от одного узла к разным узлам. Оценим каждый остовный лес по следующим параметрам:

1) Проективность (ProjectnessCoef = 0, если лес не содержит непроективного дерева, 1 – в противном случае);

2) Число незаполненных валентностей у всех узлов (ValencyMissCount);

3) Число компонент связности (ConnectedComponentsCount);

4) Число отношений, которые не прошли проверку по семантическим характеристикам (SemFetDisagreeCount);

5) Число актантов, порядковый номер которых в тексте отличается от порядкового номера в словарной статье (ValencyDisorder);

6) Число отношений, которые идут в противоположном предпочтительному направлении (DirectDisagree);

7) Общее число валентностей, которые заполнены одним из значений стандартной лексической функции (поле ЛХ) (LexFetAgreeCount)

- 8) Число узлов, построенных на лексических функциях (LexFuncsNodesCount)
- 9) Число узлов, нарушающих грамматические ограничения, записанные в поле ОГРН (GramRestrViolationsCount);
- 10) Удовлетворяет ли корень дерева морфологическим критериям (TopAgreeWithSyntaxCriteria);
- 11) Число узлов, которые, подчиняя оборот типа *друг друга*, не имеют другого актанта во множественном числе (EachOtherOborotViolationCount);
- 12) Число “плохих” узлов типа МНА. “Хороший” узел МНА – это тот, в который входит больше одного элемента, и между двумя соседними членами которого стоит запятая или союз. (MNAViolationsCount);
- 13) Число отношений, которые не согласованы с частью речи узла, в который они входят (SemRelPOSViolationsCount);
- 14) Общая длина всех отношений (RelationsLength). Длина одного отношения равна расстоянию между ближайшими словами узлов, которое оно соединяет;
- 15) Проверка согласования подлежащего и сказуемого (SubjectPredicateViolationsCount);
- 16) Число связей в графе, которые были получены не из словарных статей (число слабых связей) (RelationsNotFromRossCount);
- 17) Число ошибок, связанных с выражением МНА, взятого из словарной статьи (ValencyMNAViolationsCount);

Вес остоного леса вычисляется по следующей формуле:

$$\begin{aligned} \text{WEIGHT}(T_i) = & \text{ValencyDisorder} * 10 + \text{ValencyMiss} * 10 \\ & + \text{DirectDisagree} * 10 + \text{SemFetDisagree} * 20 + \text{ConnectedComponentsCount} * 100 + \text{ProjectnessCoef} * 80 \\ & - \text{LexFetAgreeCount} * 20 + \text{LexFuncsNodesCount} * 30 + \text{GramRestrViolationsCount} * 100 + \\ & \text{TopAgreeWithSyntaxCriteria} * 10 + \text{EachOtherOborotViolationCount} * 50 + \end{aligned}$$

$$\text{MNAViolationsCount} * 50 + \text{SemRelPOSViolationsCount} * 5 +$$

$$\text{RelationsLength} * 2 + \text{SubjectPredicateViolationsCount} * 50 + \text{RelationsNotFromRossCount} * 2 + \text{ValencyMNAViolationsCount} * 30.$$

Соответственно, чем больше вес остоного дерева, тем перспективней выбор его для дальнейшего лингвистического анализа.

Ниже мы опишем процедуры получения некоторых параметров, использованных для вычисления веса остоного дерева.

VI. Проверка отношений по семантическим характеристикам

Эта процедура проверяет отношение между узлами, главные слова которых были найдены в словаре.

Пусть Host - словарная статья главного слова узла, от которого идет рассматриваемое семантическое отношение (далее просто отношение)). Пусть Slave – словарная статья главного слова узла, к которому идет отношение.

Сейчас нам понадобятся только поля CX из этих двух словарных статей. Напомним, что поля CX содержат дизъюнкцию конъюнкций, т.е. нормальную форму, семантических характеристик или отношений. Дизъюнкты нумеруются начиная с 1. Например:

$$\text{CX}(\text{Host}) = \begin{matrix} 1 \text{ ФИН, ОРГ} \\ 2 \text{ ОДУШ} \end{matrix}$$

Пусть отношение, которое мы проверяем, стоит i-ым по порядку в словарной статье слова Host.

Считается, что отношение удовлетворяет критерию семантических характеристик, если какой-либо из дизъюнктов $\text{CX}_i(\text{Host})$ полностью вкладывается в какой-либо дизъюнкт $\text{CX}(\text{Slave})$.

Например, если:

$CX_i(\text{Host}) = \text{ОРГ}$

$CX(\text{Slave}) = \text{ОРГ, ФИН.}$

Т.е., в $CX(\text{Slave})$ записано, что Slave – финансовая организация, а в ограничениях по i -ой валентности записано, что i -ая валентность – просто организация, значит, она может быть и финансовой.

VIII. Вычисление валентной структуры для узлов, построенных на лексических функциях

Рассмотрим случаи взаимодействия валентных структур присвязочного глагола и слова-ситуации.

Если присвязочный глагол не был найден в словаре, то валентная структура вычисляется в соответствии с определением лексических функций, данным в классических работах Московской лингвистической школы (см. [6]). Например, если X – узел, построенный на функции Орег_1 , то первая валентность этого узла синтаксически выражается подлежащим, а семантические ограничения копируются из первой валентности слова-ситуации. Вторая валентность узла X – само слово-ситуация. Все остальные валентности узла X берутся опять же из слова-ситуации, в том порядке, в котором они там перечислены.

Если присвязочный глагол есть в словаре, то нужно решить, как соединять валентные структуры присвязочного глагола и слова-ситуации. Валентности присвязочного глагола может заполнять очень широкий (по семантическим критериям) класс узлов, что отражается на размере полей CX_i . Грамматические ограничения валентностей присвязочного глагола, наоборот, предельно конкретны.

В каком-то смысле слова-ситуации являются антиподами присвязочных глаголов. Слова-ситуации обладают слабыми грамматическими средствами управления и богатой семантической структурой. В предельном случае слово-ситуация вообще не имеет синтаксических средств выражения некоторых своих семантических валентностей. Но почти всегда существует опосредованный способ – в соединении с каким-нибудь присвязочным глаголом, который даст грамматическое оформление валентности. Например, у слова *экзамен* есть семантическая валентность на субъекта ситуации экзамена (тот, кто принимает экзамен), но активной грамматической такой же валентности нет.

Таким образом, правильно было бы дать узлу X (построенному на лексической функции) семантические валентности от слова-ситуации, а грамматические – от присвязочного глагола.

Однако в РОССе почти нет валентностей, для которых не существует грамматического выражения, поэтому все это построение неприменимо.

Единственное, что нам остается, – это расширить грамматические характеристики валентностей слова-ситуации грамматическими характеристиками присвязочного глагола.

Литература

[1] Н.Н.Леонтьева. Строение семантического компонента в информационной модели автоматического понимания текста. Автореф. и дисс. д.т.н. М., 1990.

Н.Н.Леонтьева. ПОЛИТекст": информационный анализ политических текстов. // НТИ, сер.2, N 4, 1995. – С. 3-21.

Н.Н. Леонтьева. Русский общесемантический словарь (РОСС): структура, наполнение. // НТИ. Сер. 2. - 1997. - N 12. - С.5-20.

Семенова С.Ю. Прилагательные в семантическом словаре

одной прикладной системы. // Труды Международного семинара

Диалог'98 по компьютерной лингвистике и ее приложениям. -

С.553-564.

Кулагина О.С. Исследования по машинному переводу. М., 1979.

Мельчук И.А. Опыт теории лингвистических моделей "СмыслоТекст".Семантика, синтаксис. – М.:Наука, 1974.

Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. М.,1985 г.

Сокирко А.В. Исследование слов с характеристиками

'информация' и 'носитель информации' в русском общесемантическом

словаре. // Труды Международного семинара

Диалог'98 по компьютерной лингвистике и ее приложениям.

[1] В программной реализации все несколько сложнее, поскольку полученный на предыдущем этапе граф может быть не связан.

[2] Остовное дерево связного графа – это максимальный подграф, который является деревом. Очевидно, что остовных деревьев у одного связного графа может быть много. Мы можем сказать только, что число остовных деревьев полного графа с числом вершин n больше $n!$.