Меркурьев Дмитрий Васильевич

ОБРАБОТКА СУЩЕСТВИТЕЛЬНЫХ В СИНТАКТИКО-СЕМАНТИЧЕСКОМ АНАЛИЗАТОРЕ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

1. ВВЕДЕНИЕ

Информационные системы поиска, классификации, аннотирования и перевода необходимы для эффективной работы с большими коллекциями текстов на естественных языках. Добиться существенного улучшения функциональности данных систем можно посредством использования методов, основанных на анализе семантики документов. Функциональная теория языка, разработанная Тузовым В.А., является успешным подходом к решению задачи определения смысла текста [1, 2]. На основании данной теории были разработаны синтактико-семантические анализаторы, которые способны не только получить синтаксическую структуру предложения, совпадающую с семантической, но и определить смысл каждого слова предложения, выражаемый толкованием на семантическом языке.

Обработка существительных является важным этапом работы синтактико-семантического анализатора и заслуживает отдельной статьи, так как качество анализа текста существенно зависит от правильного выбора семантической альтернативы слов данной части речи.

2. КРАТКИЙ ОБЗОР СУЩЕСТВУЮЩИХ СИСТЕМ

На данном этапе существует несколько систем, которые претендуют на синтаксический или частичный семантический анализ текстов на русском языке. Синтаксический анализатор «DictaScope» [3] компании «Dictum» строит дерево подчинительных связей предложения и определяет грамматические характеристики слов. Система проекта АОТ («Автоматическая обработка текста») [4] осуществляет первичный семантический анализ текста. Также следует упомянуть систему «TextAnalyst» [5] и продукты компании RCO [6, 7].

Все вышеперечисленные системы не получают полного семантического описания текста и имеют ограничения, связанные с выбором правильной альтернативы слова в предложении. Данные обстоятельства определяют уникальность синтактико-семантических анализаторов, построенных на основе теории Тузова В.А.

3. НЕКОТОРЫЕ ПОЛОЖЕНИЯ ТЕОРИИ ТУЗОВА В.А.

Теория Тузова В.А. заслуживает самого пристального рассмотрения, так как предлагает перспективный подход к решению задач формализации и семантического анализа текстов на естественных языках, более подробно ознакомиться с которым можно в книге [2]. В данной статье будут рассмотрены лишь некоторые положения теории, необходимые для понимания специфики предметной области.

3.1. ВАЖНЫЕ ТЕЗИСЫ О ФОРМАЛЬНОЙ МОДЕЛИ ЯЗЫКА

Тезис 1.

Язык является алгебрарической системой {f1, f2, ... , fn, M}, где fi – *базисные функции*, M – структура данных (*базисные понятия*).

Тезис 2.

Каждое слово языка является именем функции, которая описывает его семантику. Предложение является суперпозицией данных функций.

Тезис 3.

Грамматика неразрывно связана с семантикой языка и может быть представлена в семантическом словаре.

3.2. БАЗИСНЫЕ ПОНЯТИЯ И ФУНКЦИИ, ОПИСАНИЕ СЕМАНТИКИ СЛОВА

Базисными понятиями являются слова естественного языка, смысл которых не выражается через более простые понятия. Классификатор организует базисные понятия в иерархическую систему классов по типу род – вид, отношения в которой удовлетворяют следующим принципам:

1) все слова данного класса имеют общие семантические свойства;

2) наследующий класс имеет все свойства наследуемого класса, а также специфические семантические характеристики. Ниже приведены примеры в следующей нотации:

```
Класс $<код класса> < имя класса >
< слова класса >

Класс $143 Знания Литература
ЛИТЕРАТУРА, СЛОВЕСНОСТЬ, ...

Класс $1430 Знания Литература Книга
АТЛАС, БРОШЮРА, БУКЛЕТ, ...

Класс $14302 Знания Литература Книга Пособие
БУКВАРЬ, ЗАДАЧНИК, ПОСОБИЕ, УЧЕБНИК, ...
```

Базисные функции являются частью семантического языка и описывают отношения между аргументами. Ниже приведены некоторые базисные функции:

- **And(x, y)** – x и y
- **Caus(x, y)** – x является причиной y
- **Control(x, y)** – x управляет y
- **Func(x)** – имеет место быть x
- **Hab(x, y)** – x имеет y
- **Incep(x)** – x начинается
- **Lab(x, y)** – x подвергается действию y
- **Loc(x, y)** – x находится в y
- **Oper(x, y)** – x совершает y
- **Rel(x, y)** – x имеет отношение к y
- **Perf(x)** – x завершился

При помощи суперпозиции базисных понятий и функций выражается семантика производных понятий:

```
\\ "нечто, имеющее отношение к литературе"
ЛИТЕРАТУРНЫЙ   N%~ЛИТЕРАТУРА$143
(A1>Rel(A1:НЕЧТО$1,ЛИТЕРАТУРА$143))
\\ "книга, имеющая три тома "
ТРЕХТОМНИК  $1430(S1>Hab(S1:КНИГА$1430,
ТРИ$12/031(ТОМ$14303)))
```

3.3. СЕМАНТИКО-ГРАММАТИЧЕСКИЕ ТИПЫ, АРГУМЕНТЫ И ГРАММАТИЧЕСКИЕ ЗНАЧЕНИЯ АЛЬТЕРНАТИВ СЛОВА, ДОПОЛНИТЕЛЬНЫЕ АРГУМЕНТЫ

Семантико-грамматический тип может состоять из комбинаций следующей информации: семантического класса, унификатора рода и числа, предложно-падежной формы (**$1223/02~@ОНЪ$17@Им**, **$12/1~!Где**, **$1223/02~@ОНЪ$17@Вин**, **$12/1~!наПред**, **@ОНЪ$17@Им**, **@Род**, **@Куда**). Обобщенные семантические значения обозначаются следующими лексемами: **@Куда**, **@Откуда**, **@Когда**, **@Где**, и т. д.

Семантическая альтернатива слова может быть представлена описанием следующего вида:

```
<основная форма слова> { < морфологическая
информация > СГТ11, СГТ12,…, СГТ1n }
< семантический класс > ( Z1: СГТ21, Z2:
СГТ22,…, Zm: СГТ2m),
```

где **СГТij** – семантико-грамматические типы, **Zi** – обозначения переменных.

Следует отметить, что в случае производных слов каждому такому описанию можно сопоставить еще и толкование смысла альтернативы на семантическом языке в виде суперпозиции базисных функций и понятий. В фигурных скобках записываются *семантико-грамматические значения* альтернативы, а в круглых – ее *аргументы*. Первые содержат информацию о том, к каким альтернативам других слов данная альтернатива может присоединиться как значение аргумента, вторые – какие альтернативы она может присоединить. Например, в процессе обработки словосочетания «в лесу» синтактико-семантический анализатор выберет 196 альтернативу предлога «в» и третью – слова «лесу».

Выбранная альтернатива первого слова через аргумент **$122412~!Пред** присоединит соответствующую альтернативу второго слова, и для собранной конструкции будет выработан семантико-грамматический тип **$122412~@Где**:

```
@Где в<196>
    (@Пред лесу<003>)
```

// <196>, <003> – номера альтернатив.

Выбранные альтернативы:

```
<196> В {Предлог. $122412~@Где} (Z0: y>
@Где, Z1: ПОЛЕ$1224113 \ ПАРК$122412
\ ПЕРСПЕКТИВА$12/1200 ~! Пред)

<003> ЛЕС {Сущв._Муж_Неодуш
$122412~@ОНЪ$17@Пред} $122412()
```

В зависимости от класса и части речи, на этапе предварительной обработки (до синтактико-семантического анализа) к альтернативам приписываются дополнительные аргументы, которые не содержатся в первоначальном словаре: **!0Какой**, **!Вместе**, **$1210/08~@КакВ**, **!Где**, **!Когда**, **!Как**, **!Почему**, **!Зачем** и т. д. Кроме того, некоторые из приписываемых аргументов обозначаются неактивными: **хприГде**, **хОткуда**, **хКуда**, и др. Если предложение собирается не полностью, то данные аргументы становятся доступными для использования в процессе анализа.

3.4. СЕМАНТИЧЕСКИЙ СЛОВАРЬ

Основной словарь содержит описание семантики более чем 110000 слов русского языка. Кроме толкования слов на семантическом языке, данный словарь содержит синтактико-семантическую информацию, которую использует анализатор. Благодаря модулю морфологического анализа и пословной обработки, общее число обрабатываемых словоформ превышает 2250000.

4. ОСНОВНЫЕ ПРИНЦИПЫ РАБОТЫ СИНТАКТИКО-СЕМАНТИЧЕСКОГО АНАЛИЗАТОРА

4.1. ЭТАПЫ АНАЛИЗА

Работа синтактико-семантического анализатора состоит из следующих шагов:

1) морфологического анализа;

2) пословной обработки (вычисление семантико-грамматических типов, формирование независимых альтернатив слов в формате, пригодном для синтактико-семантического анализа);

3) синтактико-семантического анализа.

На третьем шаге выбираются правильные семантические альтернативы слов, которые собираются в единую конструкцию. Основными операциями на данном этапе являются нахождение связей между словами и сборка связанных альтернатив в конструкцию, при которой одна альтернатива упрятывается под другую.

4.2. НАХОЖДЕНИЕ СВЯЗЕЙ МЕЖДУ СЛОВАМИ

После морфологического анализа и пословной обработки описание предложения приобретает следующий вид:

```
< слово 1 >
          < семантическая альтернатива 1>
          < семантическая альтернатива 2>
          ...
          < семантическая альтернатива n1>
< слово 2 >
          < семантическая альтернатива 1>
          < семантическая альтернатива 2>
          ...
          < семантическая альтернатива n2>
...
< слово n >
          < семантическая альтернатива 1>
          < семантическая альтернатива 2>
          ...
          < семантическая альтернатива nn>,
```

где **< семантическая альтернатива> ::= < < номер альтернативы > основная форма слова { морфологическая информация, семантико-грамматические значения } (семантико-грамматические типы аргументов) << дополнительные аргументы >> >**.

В качестве примера далее приведены 3 альтернативы слова лист в описанной выше нотации:

```
лист

<001> ЛИСТ {Сущв. Муж Неодуш
$1213115~@ОНЪ$17@Им $1213115~@ОНЪ$17@Вин}
$1213115(Z1: !Род, Z2: !Для) << !1Какой;
!ИмС; хГде; хСравн; $1210/08~!КакВ;
```

```
!ОКакой; !уГде; !уИмея; хприГде; хприКогда; хприИмея; хсТв; хОткуда; хВключая; хБез; хДееКак; хКуда; хДля; хПротив; хКогда; хКакДолго; >>

<002> ЛИСТ {Сущв. Муж Неодуш $121316~@ОНЪ$17@Им $121316~@ОНЪ$17@Вин} $121316(Z1: !Род, Z2: !Для, Z3: !Ото) << !1Какой; !ИмС; хГде; хСравн; $1210/08~!КакВ; !ОКакой; !уГде; !уИмея; хприГде; хприКогда; хприИмея; хсТв; хОткуда; хВключая; хБез; хДееКак; хКуда; хДля; хПротив; хКогда; хКакДолго; >>

<003> ЛИСТ {Сущв. Муж Неодуш $1223/02~@ОНЪ$17@Им $1223/02~@ОНЪ$17@Вин} $1223/02(Z1: РАСТЕНИЕ$1223 ~ !Ото \ !Род) << !ИмС; хГде; хСравн; $1210/08~!КакВ; !ОКакой; !уГде; !уИмея; хприГде; хприКогда; хприИмея; хсТв; хОткуда; хВключая; хБез; хДееКак; хКуда; хДля; хПротив; хКогда; хКакДолго; >>
```

Синтактико-семантический анализатор определяет 2 основных типа связей:

1) между семантическими аргументами присоединяющего слова и семантико-грамматическими значениями присоединяемого (глагол присоединяет существительное, предлог – существительное и т. д.);

2) соответствия среди семантико-грамматических значений двух слов (согласование прилагательного или причастия с существительным и т. д.).

Ниже приведен пример обработки предложения и список полученных связей.

Пример

«На столе лежал лист бумаги».

```
@Глагол лежал<X003.006>
    (@Где На<X001.052>
        (@Пред столе<X002.001>,
    @Им лист<X004.002>
        (@Род бумаги<X005.003>))
```

Связи между словами:

```
<X001.052> $12~@Пред => <X002.001>
<X003.006> $12~@Где => <X001.052>
<X003.006> $12~@наПред => <X001.052>
<X003.006> $121316~@ОНЪ$17 => <X004.002>
<X004.002> @Род => <X005.003> ,
```

где **<X< номер слова >.< номер альтернативы >.**

Как видно из примера, при анализе словосочетания «лист бумаги» была получена связь **<X004.002> @Род => <X005.003>** и выбрана вторая альтернатива слова «лист».

4.3. СБОРКА КОНСТРУКЦИИ ИЗ ПРИСОЕДИНЯЮЩЕГО И ПРИСОЕДИНЯЕМОГО СЛОВ

Данную конструкцию можно представить в скобочной форме. В приведенном ниже примере 1-я альтернатива слова «холм» упрятывается под 72-ю предлога «на», и для структуры вырабатывается обобщенное семантико-грамматическое значение **@Где**.

```
на холме
@Где на<X001.072>
  (@Пред холме<X002.001>)
```

Следует отметить, что присоединенные слова не теряют активности и могут участвовать во взаимодействии на дальнейших этапах анализа. Данное обстоятельство позволяет анализатору просматривать предложение один раз слева направо независимо от степени вложенности подконструкций.

5. АЛГОРИТМ АНАЛИЗА ПРЕДЛОЖЕНИЯ И ОБРАБОТКА СУЩЕСТВИТЕЛЬНЫХ

На данном этапе алгоритм анализа предложения реализован в виде связки рекурсивных функций, каждая из которых обрабатывает конкретную часть речи. В зависимости от альтернатив слова, вызываются функции обработки глаголов, существительных, прилагательных, наречий, предлогов и т.д. Данные функции описывают те роли, которые соответствующие части речи могут играть в предложении. Ниже приведен пример работы алгоритма анализа предложения.

Пример

«На поляне выросло дерево».

Процесс анализа (см. распечатку 1).

Распечатка 1

```
обработка предложения
      обработка группы предлога
            На \ поляне
        обработка группы существительного
          поляне \ выросло
             \выход
                  обработка_предлога
          На \ поляне
                  \выход СВЯЗЬ: предлог и существительное: На => поляне
  \выход СВЯЗЬ: глагол и группа предлога: выросло => На
  обработка глагола
            выросло \ дерево
                  обработка группы существительного
                        дерево \ дерево
      \выход
      \выход СВЯЗЬ: глагол и существительное: выросло => дерево
\выход
```

Результаты анализа:

Скобочная структура:

```
@Глагол выросло<X003.002>
// глагол и группа предлога
    (@Где На<X001.090>
// предлог и существительное
      (@Пред поляне<X002.001>),
// глагол и существительное
    @Им дерево<X004.002>)
```

Смысл слов и связи:

```
На
// выбранная альтернатива с синтактико-
семантической информацией:
<X001.090> НА {Предлог. $122412~@Где}
(Z0:у> @Где, Z1: ПРИРОДА$122
\ ГРАНИЦА$12/15/16 \ РАССТОЯНИЕ$12/32
\ ПЛОЩАДЬ$12316 ~! Пред)
// семантика:
<X001.090> НА Y1>Loc(Y1:, ПРЕД:Z1)
// связь:
Z1: $122~@Пред => <X002.001>
// связь:
Z0:  у> @Где <= <X003.002>

поляне
<X002.001> ПОЛЯНА {Сущв._Жен_Неодуш
$122412~@ОНА$17@Пред} $122412(Z1:
НЕЧТО$1~!Род, Z2: РАСТЕНИЕ$1223~!Род)
<X002.001> ПОЛЯНА (РОД:Z1, РОД:Z2)
Z1: $122~@Пред <= <X001.090>

выросло
<X003.002> ВЫРАСТИ {Глагол._Сред $100/
4~@Глагол} N%~РОСТ$100/4(Perf Z1:
ДУБ$12231 \ НЕЧТО$1 ~ !ОНО$17 \ !Я$17
\ !ТЫ$17, Z2: НЕЧТО$1 ~! До, Z3: !Изо
\ !Откуда, Z4: !Тв \ !наВин)
<X003.002> ВЫРАСТИ PerfOper01(Z1,
РОСТ$100/4(ДО:Z2, ИЗО:ОТКУДА:Z3, ТВ:Z4))
@Где => <X001.090>
Z1: $12231~@ОНО$17 => <X004.002>

дерево
<X004.002> ДЕРЕВО {Сущв._Сред_Неодуш
$12231~@ОНО$17@Им} $12231(Z1: !Род, Z2:
НЕЧТО$1 ~ !Откуда)
<X004.002> ДЕРЕВО (РОД:Z1, ОТКУДА:Z2)
Z1: $12231~@ОНО$17 <= <X003.002>
```

6. ПРИМЕРЫ ОБРАБОТКИ СУЩЕСТВИТЕЛЬНЫХ

Как уже писалось выше, в процессе анализа предложения во взаимодействие вступают объекты, которые могут быть или отдельными словами, или уже собранными конструкциями. Условия определения типа взаимодействия обычно записывается в форме, подобной данной: если текущий объект является (содержит альтернативы) одной частью речи, а следующий – другой, и между ними есть соответствующая связь, то либо первый объект присоединяет второй, либо наоборот. Оба объекта могут быть как одной, так и разными частями речи. Кроме того, в условиях часто проверяется не только принадлежность объекта к части речи, но и то, каким членом предложения он является, следует ли

обрабатывать данное слово как особый случай. Синтактико-семантический анализатор распознает очень большое число вариантов взаимодействия данных объектов. Ниже приведены лишь некоторые из ситуаций, в которых участвуют объекты, содержащие альтернативы существительных.

Пример 1. Глагол и существительное

«Представители согласовывали условия проекта» // локальный контекст предложения

```
@Глагол согласовывали<001>
// грамматический тип, слово, номер альтернативы
    (@Им Представители<001>)
// присоединяемое слово
```

Связь: **<001> $1241~@ОНИ$17 <001>**

```
// номер альтернативы присоединяющего слова, тип связи, номер альтернативы присоединяемого слова
```

Пример 2. Инфинитив и существительное

«исследовать при помощи»

```
@Инфин исследовать<001>
    (@ДеепКак при_помощи<011>)
```

Связь: **<001> @ДеепКак <011>**

Пример 3. Два существительных взаимодействует друг с другом

«совершенствование словаря»

```
@Им совершенствование<001>
    (@Род словаря<001>)
```

Связь: <001> @Род <001>

Пример 4. Однородные существительные

«разработки, использования»

```
@Род разработки<002>
    (@Род использования<001>)
```

Связь: **<002> Однородные существительные <001>**

Пример 5. Предлог и существительное

«в парке»

```
@Где в<196>
    (@Пред парке<003>)
```

Связь: **<196> $122412~@Пред <003>)**

Пример 6. Прилагательное и существительное

«международные связи»

```
@Им связи<001>
    (@Им международные<002>)
```

Связь: **<001> прилагательное и существительное <002>**

Пример 7. Существительное и наречие

«лидер вместе со своей командой»

```
@Им лидер<002>
    (@Вместе вместе_со<004>)
```

Связь: **<002> @Вместе <004>**

Пример 8. Существительное и слово «который»

«анализ, результаты которого»

```
@Им результаты<001>
    (@Род которого<002>)
```

Связь: **<001> @Род <002>**

Пример 9. Прилагательное и предложная группа

«Подписанного на форуме»

```
@Вин подписанного<X001.002>
    (@Где на<X002.052>\<X002.097>
        (@Пред форуме<X003.002>< >)) ,
```

где **<X001.002>** – номер слова, номер альтернативы.

Связи:

```
<X001.002> $1~@Где <X002.052>
// прилагательное и обстоятельство
<X001.002> $1~@Где <X002.097>
// прилагательное и обстоятельство
<X002.052> $1241100~@Пред => <X003.002>
// предлог и существительное
<X002.097> $1241100~@Пред => <X003.002>
// предлог и существительное
```

Пример 10. Существительное и обстоятельство

«Включение в состав»

```
@Им Включение<X001.002>
    (@Куда в<X002.214>
        (@Вин состав<X003.001>))
```

Связи:

```
<X001.002> $1~@Куда <X002.214>
// существительное и обстоятельство
<X002.214> $12/113~@Вин <X003.001>
// предлог и существительное
```

Пример 11. Местоимение-прилагательное и существительное

«свои методы»

```
@Им методы<002>
    (@Им свои<001>)
```

Связь: **<002> местоимение-прилагательное и существительное <001>**

Пример 12. Существительное и причастный оборот

«Алгоритм, реализующий данный подход»

```
@Им Алгоритм<X001.001>
    (@Им реализующий<X002.001>
        (@Вин подход<X004.002>
            (@Вин данный<X003.006>)))
```

Связи:

```
<X001.001> существительное и причастный
оборот <X002.001>
<X002.001> @Вин <X004.002>
// прилагательное и существительное
<X004.002> существительное и слово данный
<X003.006>
```

7. ЗАКЛЮЧЕНИЕ

Анализаторы, построенные на основе теории Тузова В.А., являются уникальными системами, способными осуществлять полный синтактико-семантический анализ текстов на русском языке. Данные системы могут анализировать не только газетно-журнальные, но и художественные тексты.

Обработка существительных является важной частью синтактико-семантического анализатора, от которой зависит качество анализа текста. Результаты тестирования говорят о точности работы системы, составляющей от 95% до 100%, в зависимости от сложности синтаксической структуры текста. Кроме того, анализатор удобен для модификации, добавления обработчиков новых ситуаций и коррекции ошибок.

Существуют возможности адаптации анализатора для работы с другими естественными языками, что потребует разработки новых модулей предварительной обработки текста (морфологический анализ, получение необходимой для синтактико-семантического анализа информации). В случае обработки морфологии, возможно использование уже готовых систем с незначительной адаптацией. Кроме того, ядро синтактико-семантического анализатора и основная часть словаря, описывающего семантику слов, являются универсальными, что существенно упростит процесс разработки систем анализа других языков.

К сфере применения анализатора относятся следующие системы: поисковые, вопросно-ответные, классификации документов, аннотирования, реферирования, фильтрации спама, перевода, пополнения баз знаний.

Литература

1. *Тузов В.А.* Математическая модель языка. Л.: Изд-во Ленингр. ун-та, 1984. 76 с.
2. *Тузов В.А.* Компьютерная семантика русского языка. СПб.: Изд-во СПбГУ, 2004. 400 с.
3. Вебсайт проекта: http://www.dictum.ru/
4. Вебсайт проекта: http://www.aot.ru/
5. Вебсайт компании: http://www.analyst.ru/
6. *Ермаков А. Е.* Этапы лингвистического анализа текста в программных продуктах RCO – Русский язык: исторические судьбы и современность. II Международный конгресс исследователей русского языка. Труды и материалы. Москва, МГУ, 2004.
7. Вебсайт компании: http://www.rco.ru/

Меркурьев Дмитрий Васильевич,
аспирант кафедры информатики математико-механического факультета СПбГУ.