МАКСИМОВ Андрей Иванович -
кандидат технических наук, доцент кафедры электронных документов, архивов и технологий Российского государственного гуманитарного университета (РГГУ), старший научный сотрудник Лаборатории автоматизированных систем управления НИИ СП им. Н.В. Склифосовского
Адрес: 125993, ГСП-3, г. Москва, Миусская пл., 6
e-mail: a.i.maks@mail.ru

МОДЕЛИ И МЕТОДЫ КОМПЬЮТЕРНОЙ СЕМАНТИКИ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

Введение

Значительная часть информационных ресурсов современного общества, обращающихся в компьютерных средах, представлена естественно-языковыми текстами (ЕЯ-текстами). Ввиду стремительного роста объемов такой информации все большую актуальность приобретают вопросы автоматической обработки текстов (АОТ), затрагивающие обширный спектр практических приложений, представленных задачами извлечения знаний из текстов (Text Mining), задачами автоматической классификации и кластеризации текстовых документов, автоматического индексирования и реферирования, задачами семантического поиска и множеством других задач.

Информационной основой всех перечисленных задач являются задачи выявления тематически значимой (актуальной) информации, содержащейся в анализируемом тексте, и ее идентификации посредством соотнесения с тем или иным общепринятым понятием данной предметной области.

Совокупность таких понятий можно рассматривать как некоторую понятийную спецификацию предметной области, наиболее универсальным представлением которой может служить некоторая онтология предметной области (ОПО). Таким образом, указанные выше задачи АОТ в той или иной мере сводятся к задаче определения тематически значимых слов и словосочетаний из ЕЯ-текстов и приведения их к одному из понятий ОПО. Это, в свою очередь, предполагает использование некоторой компьютерной технологии семантического анализа текстов.

Проблемы моделирования языковой деятельности человека и примыкающие к ним задачи семантического анализа естественно-языковых текстов с давних пор находятся в фокусе приоритетных исследований по искусственному интеллекту и компьютерной лингвистике.

В числе наиболее известных и основополагающих исследований в данной проблематике следует назвать работы Т. Винограда, Р. Шенка, Ч. Филлмора, Н. Хомского. В отечественной математической лингвистике сфера семантических исследований представлена работами И.А. Мельчука, Ю.Д. Апресяна, Н.Н. Леонтьевой, Е.В. Падучевой и ряда других ученых [1].

Одна из наиболее известных реализаций АОТ-инструментария - семейство программных продуктов TextAnalyst, формальной основой которого является аппарат искусственных нейронных сетей [2]. Другой, менее известный, но достаточно интересный в теоретическом и практическом аспектах подход представлен в технологии семантического анализа В.А. Тузова.

Основные положения компьютерной семантики В.А. Тузова

В основание компьютерной семантики В.А. Тузова положена функциональная модель языка. Основные теоретические положения данного подхода заключаются в том, что *1) язык представляет собой алгебраическую систему* $\{f_1, f_2, \ldots f_n, M\}$, *где* f_i *- базисные функции на языке, а M - структура языка, представляющая собой набор базисных понятий* $m_1, \ldots m_r$ *и их иерархию; 2) каждое предложение языка можно представить в виде суперпозиции базисных функций* f_i, *через которые выражается и каждое слово языка за исключением базисных понятий* $m_j \subset M$, *т.е. предложение представляет собой единую, законченную суперпозицию функций и трактуется как выражение (в математическом смысле этого слова); 3) грамматика неразрывно связана с семантикой языка, основой которой является семантический словарь, описывающий более ста тысяч лексических единиц (слов и словосочетаний), и каждое слово описывается в виде семантической формулы, состоящей из базисных функций* [3].

Наиболее существенной особенностью указанного подхода, отличающей его от других существующих подходов к формализации языка, является положение о компьютерном толковании смысла слова на формальном семантическом языке и функциональное использование этого толкования при определении смысла предложения. В семантическом языке используется около пятидесяти базисных функций, некоторые из которых представлены в **таблице 1.**

Таблица 1

Некоторые базисные функции семантического языка

Функция	Описание функции
Caus (X,Y)	X является причиной Y
Lab (X,Y)	X подвергается воздействию Y
Loc (X,Y)	X находится в Y
Hab (X,Y)	X имеет Y
Usor (X,Y)	X используется для Y

Семантический словарь содержит более ста тысяч лексических единиц, разбитых на 3 семантических уровня: *фундаментальный,* состоящий из 1500 иерархических классов и набора базисных функций; *вариативный,* состоящий из 23000 классов, которые тесно связаны с фундаментальным уровнем, они являются вариациями фундаментальных понятий и описываются на основе понятий фундаментального уровня; *описательный,* в котором слова и понятия, имеющие смысл, выходящий за рамки фундаментального и вариативного, описываются на основе понятий первого и второго уровней. Каждая словарная статья содержит заголовочное слово и его толкование на семантическом языке. Многие слова содержат несколько альтернативных толкований [3].

Результаты семантического анализа и их применение в задаче идентификации понятий

Некоторые прикладные аспекты данной технологии семантического анализа удобно показать на конкретном примере.

Рассмотрим пример семантического анализа следующего предложения:

Американское космическое агентство выделило 270 миллионов долларов четырем частным компаниям, занимающимся разработкой нового космического транспорта.

Приведенные ниже результаты семантического анализа данного предложения представлены двумя блоками.

выделило<X004.002>
(@Им агентство<X003.003>
(@Им Американское<X001.003>,
@Им космическое<X002.002>),
@Вин миллионов<X006.001>
(@Им 270<X005.001>,
@Род долларов<X007.002>),
@Дат компаниям<X010.003>
(@Дат частным<X009.004>
(@Дат четырем<X008.001>),
@Дат занимающимся[КОМПАНИЯ<X010.003>
]<X012.004>
(@Тв разработкой<X013.001>
(@Род транспорта<X016.001>\<X016.002>
(@Род нового<X014.001>,
@Род космического<X015.002>))))).

Первый блок результатов представляет собой дерево синтаксического разбора предложения, каждый элемент которого отмечен специальной маркировкой, помещенной между символами “< >” и состоящей из двух кодовых значений, разделенных точкой.

Первое кодовое значение определяет порядковый номер слова в анализируемом предложении, второе значение - номер выбранной в процессе семантического анализа семантической альтернативы данного слова (лексемы) в семантическом словаре. Данный блок результатов анализа рассматриваемого предложения для наглядности удобно представить в виде графа **(рис. 1)**, текстовым представлением которого по сути и является приведенная выше структурированная запись.

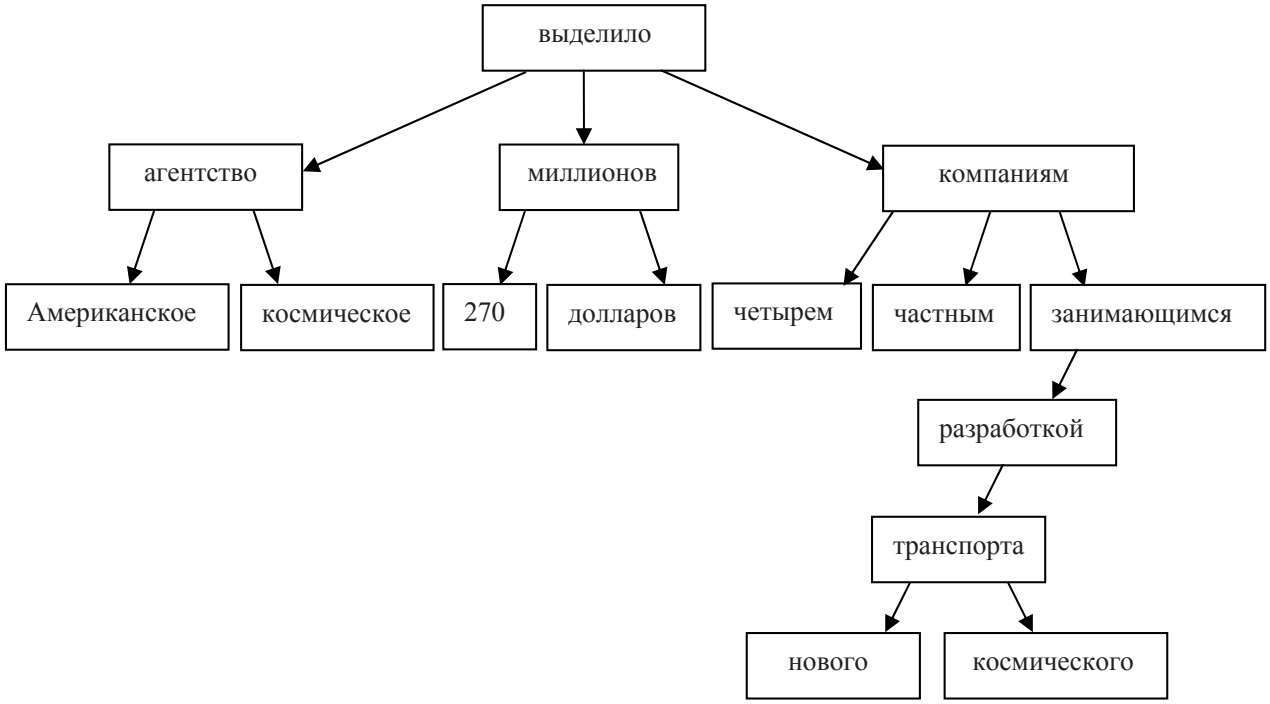

Рис. 1. *Граф синтаксического разбора предложения*

Во второй блок результатов входят формализованные описания слов анализируемого предложения, выбранные из семантического словаря. Ниже выборочно представлены значения некоторых описаний.

Агентство
АГЕНТСТВО {Сущв._Сред_Неодуш $1234091~@ОНО$17@Им}
$1234091(Z1:ЧЕЛОВЕК$1241\ПОСЕЛЕНИЕ $123~!Род,Z2: НЕЧТО$1~!поДат)
АГЕНТСТВО (РОД:Z1,ПОДАТ:Z2)

выделило
ВЫДЕЛИТЬ {Глагол._Сред}
N%~ВЫДЕЛЕНИЕ$15013(Perf Z1: !ОНО$17\ !Я$17\!ТЫ$17,Z2: !Дат\!Для\!наВин,Z3: !Вин)
PerfCaus(Z1,IncepHab(ДАТ:Z2,РОД:Z3))

долларов
ДОЛЛАР {Сущв._Муж_Неодуш $12141~@ОНИ$17@Род}
$12141(Z1: !Род,Z2: НЕЧТО$1~!Куда\!заВин\ !наВин,Z3: !Сколько)
ДОЛЛАР (РОД:Z1,КУДА:ЗАВИН:НАВИН: Z2,СКОЛЬКО:Z3)

компаниям
КОМПАНИЯ {Сущв._Жен_Неодуш $123411~@ОНИ$17@Дат}
$123411(Z0:s> ЛЮДИ$12411,Z1: СТРАНА$1231\ ЧЕЛОВЕК$1241~!Род,Z2: НЕЧТО$1~!Род)
Hab(S1:ЛЮДИ$12411(РОД:Z1,РОД:Z2), ДЕЛО$1523)

разработкой
РАЗРАБОТКА {Сущв._Жен_Неодуш $1521~@ОНА$17@Тв}
$1521(Z1: ЧЕЛОВЕК$1241~!Тв,Z2: НЕЧТО$ 1~!Род)
Caus(Oper01(Z1,РАБОТА$1521(ТВ:Z1)),IncepFunc (РОД:Z2))

космического
КОСМИЧЕСКИЙ {Прил._МужСр $1241/4~@ОНО$17@Род}
N%~КОСМОС$1227(Z0:a>НЕЧТО$1)
Rel(A1: НЕЧТО$1,КОСМОС$1227)

транспорта
ТРАНСПОРТ {Сущв._Муж_Неодуш $121324~@ОНЪ$17@Род}
$121324(Z1: !Род,Z2: !Для)
ТРАНСПОРТ (РОД:Z1,ДЛЯ:Z2)

Формализованное описание каждого слова содержит три основных раздела (строки). В первом разделе указываются некоторые грамматические характеристики (часть речи, род, одушевленность/неодушевленность) и семантико-грамматический тип

слова. Семантико-грамматический тип представляет собой комбинированное значение, первой составляющей которого является значение класса, к которому отнесено данное слово при семантическом анализе. Следующий раздел (строка) представляет собой описание слова с точки зрения его возможных сочетаний с другими лексическими переменными. Последняя строка содержит формализованное описание слова на семантическом языке.

Сказанное можно пояснить на примере описания слова РАЗРАБОТКА. В первой строке описания указаны грамматические характеристики (существительное, женского рода, неодушевленное), указан семантико-грамматический тип данного слова, главной семантически значимой составляющей которого является определенный в процессе анализа семантический класс данного слова - *$1521* - РАБОТА. Следующая строка содержит информацию о возможных лексических аргументах выбранной альтернативы рассматриваемого слова. В данном случае указано два возможных аргумента - Z1 и Z2. Значение первого аргумента может быть представлено некоторым существительным класса *$1241* (ЧЕЛОВЕК) в творительном падеже. Значением второго аргумента может служить любое существительное верхнего уровня семантической иерархии - *$1* (НЕЧТО) - в родительном падеже. Далее следует формализованная запись (формула) выбранной альтернативы слова РАЗРАБОТКА на семантическом языке:

Caus(Oper01(Z1,РАБОТА$1521(ТВ:Z1)),IncepFunc (РОД:Z2)).

Формула представлена суперпозицией трех семантических функций - *IncepFunc, Oper01, Caus* - и ее можно прокомментировать следующим образом:

Z1 выполняет работу, результатом которой является появление Z2.

Представленные результаты семантического анализа являются исходным материалом («полуфабрикатным сырьем») для последующей обработки данной информации, основным результатом которой должна стать идентификация тематически актуальных понятий, содержащихся в данном предложении. Например, если в составе ОПО присутствует понятие «ФИНАНСИРОВАНИЕ», то результатом обработки данных семантического разбора должна стать идентификация словосочетания *«выделило 270 миллионов долларов»* как указанного выше понятия ОПО. Процедуры обработки результатов семантического анализа предполагают наличие в компьютерной системе определенного набора правил, составляющих основу некоторой специализированной экспертной системы (ЭС), позволяющей осуществить такое соотнесение.

Наличие в результатах семантического разбора предложения информации о принадлежности слов к определенным семантическим классам, представление значений слов в виде суперпозиции базисных семантических функций, информация о возможных сочетаниях слова с другими лексическими переменными предоставляют обширный, семантически богатый материал и позволяют существенно облегчить составление подобных правил.

В качестве примера рассмотрим правило, предназначенное для идентификации смысла указанного выше словосочетания посредством соотнесения с соответствующим по смыслу понятием ОПО, построенное на основании семантической формулы данной лексической группы, узловым элементом которой является глагол ВЫДЕЛИТЬ. Данная семантическая формула имеет вид:

F = PerfCaus(Z1,IncepHab(ДАТ:Z2,РОД:Z3)).

Эту формулу можно прокомментировать следующим образом:

Z1 является причиной того, что Z2 становится обладателем Z3.

Пусть S - некоторое значение данной формулы (семантического выражения), определенное на множестве понятий ОПО. Тогда для прагматической интерпретации данной формулы, т. е. для приведения к соответствующему понятию ОПО, в базе знаний ЭС может присутствовать следующее правило:

ЕСЛИ
F = PerfCaus(Z1,IncepHab(Z2,Z3))
и Z1 = ОРГАНИЗАЦИЯ или ЧЕЛОВЕК
и Z3 = ДЕНЬГИ,
ТО
S = ФИНАНСИРОВАНИЕ

Значения переменных Z1 и Z3, подставляемые в правило при проверке его применимости к анализируемой лексической ситуации, определяются по семантическому классификатору (по старшим разрядам семантического кода из формализованного описания слова): Z1 - *АГЕНТСТВО ($1234091) - старшие разряды семантического кода - $1234 (УЧРЕЖДЕНИЕ, ОРГАНИЗАЦИЯ); Z3 - ДОЛЛАР ($12141) - старшие разряды кода - $1214 (ДЕНЬГИ).* Значение аргумента Z2 в правиле отсутствует. Это означает независимость рассматриваемой семантической интерпретации от значения данного аргумента.

Аргумент Z2 может вообще отсутствовать в предложении. Подобная лексическая ситуация

может быть представлена, например, следующим предложением:

Американское космическое агентство выделило 270 миллионов долларов на разработку нового космического транспорта.

В данном предложении второй аргумент глагола ВЫДЕЛИТЬ отсутствует, т. к. в предложении не определено, кому адресованы выделенные средства.

Для практической реализации базы правил, являющейся основным компонентом экспертной системы, осуществляющей извлечение тематически актуальных смыслов из ЕЯ-текстов и их идентификацию посредством понятий ОПО, необходимо использовать некоторый программный инструментарий, в качестве которого можно воспользоваться одной из оболочек экспертных систем.

Так, например, достаточно удобным подобным программным средством является среда CLIPS, представляющая собой свободно распространяемый инструментарий разработки экспертных систем [4]. База правил и программный код ЭС, разработанные в среде CLIPS, могут быть интегрированы в программные приложения, разработанные на языке C++, что существенно расширяет возможности применения технологии ЭС в составе сложных информационно-программных комплексов. Пример записи рассмотренного выше правила в нотации языка CLIPS может быть представлен следующим образом:

```
(defrule ФИНАНСИРОВАНИЕ  ""
(СЕМАНТИЧЕСКАЯ_ФОРМУЛА PerfCaus
[Z1,IncepHab[Z2,Z3]])
(or (Z1 ОРГАНИЗАЦИЯ)
(Z1 ЧЕЛОВЕК))
(Z3 ДЕНЬГИ)
(not (АКТУАЛЬНЫЙ_СМЫСЛ ?))
 =>
(assert (АКТУАЛЬНЫЙ_СМЫСЛ ФИНАНСИРОВАНИЕ))).
```

При записи семантической формулы в формате CLIPS-синтаксиса присутствующие в формуле символы круглых скобок заменены квадратными в связи с тем, что круглые скобки являются служебными символами языка CLIPS.

При срабатывании данного правила, что наступает при условии выполнения предпосылок (левой части) правила, в рабочую память ЭС добавляется факт АКТУАЛЬНЫЙ_СМЫСЛ со значением ФИНАНСИРОВАНИЕ.

Заключение

Сфера прикладных решений в проблематике АОТ в настоящее время представлена достаточно обширным спектром инструментальных средств. Однако следует иметь в виду, что любая из современных реализаций инструментария АОТ представляет собой, в большей или меньшей степени, некоторое частичное решение, и полное, универсальное решение - дело будущих разработок подобных средств, развитие которых в нескольких направлениях стимулирует творческий поиск.

Основой большинства существующих практических решений являются статистические подходы. Поэтому развитие аналитического направления в исследовании языковой деятельности, базирующееся на попытках теоретического осмысления системы естественного языка с применением тех или иных математических моделей и методов, является особенно актуальным.

Одной из интересных и перспективных реализаций технологии АОТ в рамках аналитического направления является компьютерная семантика В.А. Тузова.

Результаты семантического анализа, основанного на положениях компьютерной семантики В.А. Тузова, предоставляют обширный и удобный материал («семантический полуфабрикат»), который может быть успешно использован в составлении правил ЭС, осуществляющей извлечение тематически актуальных смыслов (знаний) из ЕЯ-текстов. Более того, этот материал является важнейшим условием, при котором возможна эффективная реализация подобной ЭС.

Применение для этих целей ЭС, правила которой основаны на результатах семантического анализа, является более универсальным решением задачи извлечения тематически актуальных смыслов из ЕЯ-текстов.

В отличие от решений подобных задач, основанных на использовании множества жестко заданных семантических шаблонов, данный подход является более гибким и позволяет обеспечить более высокую смысловую точность. При этом точность семантической идентификации в любой момент может быть повышена посредством расширения базы правил экспертной системы.

Литература:

1. Апресян Ю.Д. Избранные труды. Т. 1. - М.: Языки русской культуры, 1995. - 472 с.

2. Электронный ресурс http://www.analyst.ru/.

3. Тузов В.А. Компьютерная семантика русского языка. - СПб: Изд-во С.-Петерб. ун-та, 2004. - 400 с.

4. Частиков А.П., Гаврилова Т.А., Белов Д.Л. Разработка экспертных систем. Среда CLIPS. - СПб: БХБ-Петербург, 2003. - 608 с.