Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка.

Кузнецов И.П., Мацкевич А.Г. (ИПИ РАН)

igor-kuz@mtu-net.ru

Аннотация.

На базе русскоязычной системы АНАЛИТИК разработана ее англоязычная версия, обеспечивающая автоматическое извлечение значимой информации из текстовых сообщений на английском языке. Для этого разработан блок морфологического анализа слов английского языка. Адаптированы русскоязычные блоки лексического и синтактико-семантического анализа применительно к английскому языку. В результате создана двуязычная лингвистическая оболочка. Для нее разработаны англоязычные лингвистические знания, обеспечивающие выделение информационно-значимых объектов: лиц, адресов, дат, словосочетаний, глагольных форм и др. с автоматическим формированием содержательных портретов - семантических сетей.

Система ориентирована на сбор и анализ англоязычных документов с их автоматической формализацией и созданием пользовательской базы знаний - по примеру системы Криминал. На этой основе возможно решение различных аналитических задач, в том числе, ответ на запросы в свободной форме, а также заполнение полей какой-либо базы данных, или же составление отчетов, где в краткой форме излагается интересующая пользователя информация - в соответствии с его интересами.

Введение

За последнее время одной из важнейших проблем является автоматическая обработка текстов, получаемых пользователями, в том числе, текстов на английском языке. Лавинообразный рост объемов документов требует дифференцированного извлечения только такой информации, которая может заинтересовать пользователя. Речь идет о содержательной обработке. Трудности такой обработки определяются особенностями английского языка: многозначностью многих слов (их семантика может быть установлена только в контексте), наличием больного количества умолчаний (в том числе, в словосочетаниях) наличием сложных синтаксических конструкций, неоднозначностей и др. В связи с этим, уровень формализации текстов в существующих англоязычных системах (полнотекстовых баз данных, системах на гипертекстовой основе) невысок, что зачастую не устраивает пользователя.

Для содержательной обработки англоязычных текстов использована разработанная в ИПИ РАН система АНАЛИТИК, основанная на технологии баз знаний (БЗ) и соответствующих методиках обработки русскоязычных текстов для решения прикладных задач. Особенность методик - в переносе сложных этапов лингвистического анализа на уровень обработки знаний, а также в наличии ограничений на выделяемые объекты и глубину семантического анализа. Система базируется на концептуально-лингвистической модели и методиках, развиваемых на протяжении последних десяти лет в ИПИРАН [1]. В настоящее время удалось адаптировать эти модели и методики к английскому языку. Это оказалось возможным, так как формы русского языка покрывают многие формы английского языка. Уровень полученных результатов сопоставим с передовыми научными исследованиями за рубежом
- системы FASTUS, CIRCUS и др. [2].

Разработанная система ориентирована на обработку текстов английского языка - объявлений, сообщений о продаже различных товаров и др. Система выделяет из текстов семантически значимую информацию: интересующие пользователя объекты, их количественные, качественные характеристики и связи. Например, это могут быть

конкретные люди, их адреса, телефоны, организации, а также производства с указанием их месторасположения, состава выпускаемой продукции, их количества, качества и т.д. Их еще называют значимыми или информационными объектами. Под связями понимаются отношения (принадлежности, родственные), участие в одном действии, время, место события.

Выделение значимых объектов осуществляется лингвистическим процессором, который состоит из оболочки, управляемой лингвистическими знаниями. Настройка на выделяемые объекты и анализируемые формы языка осуществляется путем разработки соответствующих лингвистических знаний. В качестве примера система была настроена на тексты, касающихся объявлений о продаже земельных участков с аукциона, где значимые объекты - это люди, земельные участки, адреса, организации, цены и др. Хотя возможна настройка на другие тексты и объекты. Ниже рассматриваются особенности работы блока морфологического анализа и организации лингвистических знаний англоязычной системы с примерами в упомянутой прикладной области.

Англоязычная версия системы ориентирована на обработку больших потоков текстов с выдачей пользователю (аналитику) необходимой информации в наиболее удобном для него виде. Эта система решает следующие задачи:

- автоматический ввод документов с их делением на части и лексическим анализом;
- автоматическую формализацию текстовой информации с созданием собственной базы знаний (БЗ), имеется в виду направленное извлечение знаний из англоязычных текстов с ее использованием на уровне БЗ.
- составление отчетов (рефератов), имеющих вид файлов, где в краткой форме излагается интересующая пользователя информация - в соответствии с его шаблоном.

Проверка работы системы осуществлялась на примере выявления из сети ИНТЕРНЕТ информации, касающейся объявлений о продаже земельных участков с аукциона, с последующим выделением информационных (значимых) объектов по схеме, заданной пользователем: ШТАТ - ВРЕМЯ ПОСТУПЛЕНИЯ ЗАЯВКИ - СОБСТВЕННИК - ЦЕНА - МЕСТОРАСПОЛОЖЕНИЕ - ДАТА ПРОДАЖ - СВЯЗЬ (e-mail, адрес) - ОСОБЕННОСТИ УЧАСТКА.

1. Представление знаний.

Знания (предметные и лингвистические) в БЗ системы представляются в виде структур, которые записываются в нотации семантических сетей, дополненных средствами представления событийных компонент и комплексных связей. В результате образуются расширенные семантические сети (РСС). РСС состоит из элементарных фрагментов, имеющих произвольное количество аргументных мест (но не более 200) и представляющих свойства, отношения, события, действия. Множество фрагментов - это РСС [3,4].

В простейшем случае фрагмент имеет вид N-местного предиката. Например, DATA_(7,JANUARY,2002) - это фрагмент, представляющий дату. В тоже время фрагмент - это более сложная конструкция, которая далеко выходит за рамки типовых предикатов логики 1-го и 2-го порядков.

Во-первых, в фрагментах широко используются внутрисистемные коды - это числа, к которым добавляется знак плюс (+), когда вводится новый код, или знак минус (-), когда используется уже введенный код. Например, "1+" и "1-" - есть обозначение одного и того же объекта (или отношения), а "2+" и "2-" - уже другого, и т.д. Такие числа служат для обозначения неименованных объектов, например, порождаемых самой системой. Например, в фрагментах

SUB(MAN,1+) NAME(JOHN,1-)

код 1+ и 1- представляют одного и того же человека по имени JOHN.

Во-вторых, вводится специальный код фрагмента, соответствующий всей представленной в фрагменте информации. Например, в фрагменте

ОРГ_(MORTGAGE,ELECTRONIC,REGISTRATION,SYSTEM,INC./3+) код 3+ представляет всю организацию. Эти коды могут стоять на аргументных местах других фрагментов. Например, фрагменты

```
FIO(SHALMAR,REESE,DANIEL," "/2+)
ОРГ_(MORTGAGE,ELECTRONIC,REGISTRATION,SYSTEM,INC./3+)
GIVE(2-,3-)
```

представляют, что SHALMAR REESE DANIEL (ему сопоставлен код 2+, 2-) передал (GIVE) данные организации ОРГ_(MORTGAGE,ELECTRONIC,REGISTRATION,SYSTEM,INC./3+), которой сопоставлены коды 3+, 3-. Итак, коды фрагментов необходимы для представления комплексной информации и различных видов связей.

PCC ориентированы на отображение возможности интеграции множества связанных объектов в один объект, что выражается в англоязычных текстах в виде форм с причастиями (participle) и герундиями (с окончанием ING), а также отглагольными существительными (с окончаниями TION и др.). Понятие связи рассматривается в широком смысле. Это могут быть не только отношения, но и зависимости. Связанными считаются также объекты, участвующие в одном действии. Группа связанных объектов может быть связана с другой группой, что в англоязычных текстах выражается в виде глагольных форм со словами - причастиями, герундиями, а также существительными - производными глаголов.

2. Содержательные портреты документов.

Сеть (PCC), представляющая объекты и связи какого-либо документа, образует так называемый содержательный портрет этого документа. Такие портреты необходимы для обеспечения быстрого и качественного поиска информации по значимым компонентам и связям. Приведем в качестве примера типичный текст объявлений о продажах.

Georgia, Coweta County
Under and by virtue of the Power of Sale contained in a Security Deed given by Shalamar Reese to Mortgage Electronic Registration Systems, Inc., dated January 7, 2002, recorded in Deed Book 1823, Page 221, Coweta County, Georgia Records, conveying the after-described property to secure a Note in the original principal amount of Eighty-Nine Thousand Eight Hundred Twelve and 0/100 Dollars ($89,812.00), with interest thereon as set forth therein, there will be sold at public outcry to the highest bidder for cash before the courthouse door of Coweta County, Georgia, within the legal hours of sale on the first Tuesday in April, 2003, the following described property: Exhibit УАФ All that tract or parcel of land consisting of 0.409 acres, lying and being in Land Lot 91 of the Fifth Land District of Coweta County ...

Его содержательный портрет имеет вид:

```
ДОК_(4,ENG_1.TXX,"ИНТЕРНЕТ;")
PLACE_(GEORGIA,COWETA,COUNTY/0+) 0-(4,PLACE_)
 CONTAIN(VIRTUE,OF,POWER,OF,SALE,
SECURITY,DEED/1+) 1-(4,ACT_)
FIO(SHALMAR,REESE,DANIEL," "/2+)  2-(4,FIO)
 ОРГ_(MORTGAGE,ELECTRONIC,REGISTRATION,
SYSTEM,INC./3+) 3-(4,ОРГ_)
GIVE(2-,3-/4+)  4-(4,ACT_)
DATE(/5+) 5-(4,ACT_)
DATA_(7,JANUARY,2002/6+)  6-(4,DATA_)
When(5-,6-/7+) RECORD(DEED,BOOK/8+)  8-(4,ACT_)
PLACE_(COWETA,COUNTY,GEORGIA/9+)
9-(4,PLACE_)
```

```
 NUMBER_($9812/10+)  10-(4,NUMBER_)
NUMBER_("$89,812"/11+)  11-(4,NUMBER_)
SET(FORTH/12+)  12-(4,ACT_)
SELL(THERE,PUBLIC,OUTCRY,HIGHEST,BIDDER,CA
SH/13+)  13-(4,ACT_)
DATA_(1,TUESDAY,APRIL,2003/14+)  14-(4,DATA_)
SALE_ON(14-/15+)  15-(4,ACT_)
DESCRIBE(FOLLOW,PROPERTY/16+)  16-(4,ACT_)
BE(LAND,LOT/17+)  17-(4,ACT_)
PLACE_(COWETA,DISTRICT/18+)  18-(4,PLACE_)

  ПРЕДЛ_(4,0-,1-,4-,5-,6-,8-,1823,PAGE,221,9-,RECORD,CONVEY,THE,
AFTER,DESCRIBE,PROPERTY,SECURE,NOTE,IN,ORIGINAL,PRINCIPAL,AMOUNT,
OF,80,10-,11-,WITH,INTEREST,THEREON,A,12-,FORTH,THEREIN,13-,BEFORE,
COURTHOUSE,DOOR,WITHIN,LEGAL,HOUR,OF,15-,16-,EXHIBIT,УАФ,ALL,
THIS,TRACT,OR,PARCEL,OF,LAND,CONSIST,OF,0.409,ACRE,LY,AND,17-, 91,OF,5,
LAND,18-)
```

Первый фрагмент ДОК_(4,ENG_1.TXX,"ИНТЕРНЕТ;") указывает, что содержательный портрет построен на основе файла 'ENG_1.TXX', взятого из "ИНТЕРНЕТ". Он запомнен в БЗ как документ под номером 4. Второй фрагмент PLACE_(GEORGIA,COWETA,COUNTY/0+) представляет место, где происходят действия. Добавка 0-(4,PLACE_) указывает на принадлежность этого места к документу 4. Такие фрагменты необходимы для быстрого поиска нужных фрагментов, когда в оперативной памяти (БЗ) находится множество содержательных портретов. Последний фрагмент ПРЕДЛ_(4,...) содержит коды других фрагментов и представляет порядок расположения соответствующей информации в тексте документа. По ним (заменяя коды на соответствующие группы слов) можно восстановить текст.

Такие сети представляют достаточно высокий уровень формализации текстов и удобны для обработки - с помощью инструментальных средств DECL [5].

3. Лингвистический процессор

Лингвистический процессор системы обеспечивает автоматическое построение содержательных портретов. Он включает в себя лексикографический, морфологический, терминологический и синтактико-семантический анализ.

Блок лексикографического анализа обеспечивает: автоматическое деление текста на самостоятельные части, а также определения начала и конца предложения.

Морфологический анализ имеет целью - приведение слов в каноническую форму. Каждому слову присваиваются признаки, которые делятся на три группы: лексические (слово с большой буквы, большими буквами, с точкой на конце или это отдельная буква и др.), морфологические (грамматическая категория слова, число для существительных и т.д.) и семантические (фамилия, имя, отчество и др.). Количество семантических признаков может увеличиваться - за счет специальных словарей - организаций, стран, городов и др. Само слово в нормальной форме тоже считается признаком. Морфологический анализ необходим, чтобы избавиться от различных форм написания слов, и облегчает поиск.

Терминологический анализ обеспечивает выделение терминов, а также синонимичные преобразования.

Синтактико-семантический анализ осуществляется специальными контекстными правилами (см. п.6) и служит для выделения из документа значимых компонент и связей.

4. Блок морфологического анализа.

Блок морфологического анализа управляется лингвистическими (морфологическими) знаниями, имеющими вид фрагментов РСС, и обеспечивает присвоение слову признаков, а также приведение словоформ одного и того же слова к одному виду - каноническому (для существительных - это ед. число, для глаголов - инфинитив и т.д.)

Результатом работы блока морфологического анализа является семантическая сеть (РСС), представляющая пространственную структуру текста. В ней представлены слова в нормальной форме с их признаками и указанием их последовательности. Последующая обработка сводится к преобразованию сетей на основе заданных правил.

Морфологический анализа сводится к делению слова на части: КОРЕНЬ/ОКОНЧАНИЕ или КОРЕНЬ/СУФФИКС/ОКОНЧАНИЕ. Для выделения окончаний используются фрагменты:

```
M_OKON_S("IES",3,MANY," "/1+) 1-("Y")
M_OKON_S("OES",3,MANY," "/1+) 1-("O")
M_OKON_S("AES",3,MANY," "/1+) 1-("A")
M_OKON_S("YES",3,MANY," "/1+) 1-("Y")
M_OKON_S("S",-1,MANY," ")
```

Фрагмент M_OKON_S("IES",3,MANY," "/1+) 1-("Y") указывает на необходимость отделения от слова трех последних букв. И если это "IES", то делается замена "IES" -> "Y". Например, слово FLIES заменяется на FLY. Или же просто отнимается окончание "S" и слову присваивается признак MANY (см. последний фрагмент).

Для анализа глаголов (VERB) дополнительно используются фрагменты:

```
M_OKON("ED",2,VERB,PAST_)
M_OKON("ING",3,PRICH," ")
M_OKON("IED",3,VERB,PAST_/1+) 1-("Y")
```

Отделяются окончания ED ING и слову присваиваются соответствующие признаки VERB, PAST_, PRICH, а окончание "IED" заменяется на "Y".

Для выделения суффиксов используются фрагменты вида:

```
M_SUF("ION",3,VERB,OBJ)
```

где "ION" - суффикс, 3 - сколько букв в суффиксе, VERB - часть речи без суффикса (глагол), OBJ - формируемый признак "объекта".

В канонической форме суффиксы не отделяются от корней. Они служат только для выявления части речи и присвоения слову признаков.

Другие примеры фрагментов, используемых для выявления суффиксов:

```
M_SUF("LY",2," ",ADV)  {= ADV - наречие ADVERB =}
M_SUF("OR",2,VERB,OBJ)
M_SUF("IER",3,VERB,OBJ/1+) 1-("Y")  {= добавляется "Y" =}
M_SUF("ER",2,VERB,OBJ)
M_SUF("ERY",3,VERB,OBJ) {= COOK - готовить, COOKERY - стряпня =}
M_SUF("IVE",3,VERB,ADJ)
M_SUF("IST",3,ADJ,OBJ)
M_SUF("MENT",4,VERB,OBJ)  {= Суффикс MENT преобр. VERB -> OBJ =}
M_SUF("ABLE",4," ",OBJ) ....
```

Будем называть корнем слово, оставшееся после отделения окончаний "S", "ED", "ING" и др. По корням также определяется часть речи. Для этого служат фрагменты вида:

M_ROOT("DECID","DECIDE",VERB," ")

где "DECID"- корень, "DECIDE" - слово в каноническом виде, VERB - формируемый признак (их может быть два).

С помощью таких фрагментов задаются все глаголы, заканчивающиеся на "E" и "Y". Иначе их невозможно привести к каноническому виду, так как нет единого правила, указывающего, что отделять "ED" или "D", например, DECID/ED DECIDE/D или BLOCK/ED BLOCKE/D.

Другие примеры:

```
M_ROOT("DESCRIB","DESCRIBE",VERB," ")
M_ROOT("DESIGNAT","DESIGNATE",VERB," ")
M_ROOT("DECLAR","DECLARE",VERB," ")
M_ROOT("DISCLOS","DISCLOSE",VERB," ")
.....
```

Для выявления частей речи и семантических признаков (имен, фамилий и др.) используются фрагменты вида:

```
M_WORD("OF","OF",PREP," ")  {= PREP - предлог =}
M_WORD("BETWEEN","BETWEEN",PREP," ") ....

M_WORD("I","I",PRON_1," ")     {= PRON_1 – местоимения личные =}
M_WORD("ME","I",PRON_1," ")        {= Преобразование "ME" -> "I" =}
M_WORD("MY","I",PRON_1,ADJ)  {= Формируется признак ADJ - прилагательное = ....}
M_WORD("FRIEND","FRIEND",MAN," ") {= MAN - человек =}
M_WORD("MATHER","MATHER", RELL," ") {= RELL - родственное отношение  =}
M_WORD("FATHER","FATHER",RELL," ") ...

M_WORD("ANNE","ANNE",NAME," ")              {= NAME - имя =}
M_WORD("CALLIE","CALLIE",NAME," ")
M_WORD("DANIEL","DANIEL",NAME," ") ...
```

Такие же фрагменты служат для выявления неправильных глаголов:

```
M_WORD("FIND","FIND",VERB," ")
M_WORD("FOUND","FIND",VERB,PAST_)
```

В последнем случае заменяется "FOUND" -> "FIND" и добавляются признаки глагола VERB и прошедшего времени PAST_.

Другие примеры:

```
M_WORD("DRIVE","DRIVE",VERB," ") M_WORD("DROVE","DRIVE",VERB,PAST)
M_WORD("DRIVEN","DRIVE",VERB,PAST1) M_ROOT("DRIV","DRIVE",VERB," ")
```

5. Терминологический анализ и синонимичные преобразования.

Терминологический анализ имеет целью - синонимичные преобразования, расшифровку сокращений, выделение терминов. Для этого используются фрагменты следующего вида:

```
TERMIN(<результ.слово>,<слово1>,<слово2>) или
TERMIN(<результ.слово>,<слово1>,<слово2>,<слово3>),
```

где <слово1>,... это может быть - отдельное слово, признак, а также
И-ИЛИ графы. Фрагменты типа "ИЛИ" представляется STR_OR(...), где перечисляются факультативные слова или их признаки. Фрагменты типа "И" представляется STR_AND(...), где предполагается обязательность слов с указанными признаками.

Например, TERMIN(WHERE,HOW,FAR) обеспечивает преобразование "HOW FAR" -> "WHERE".

Другой пример:

TERMIN_('P.O.','P.','O.') {= склеиваются буквы "P." и "O." =}

Более сложный случай:

TERMIN_(1,NUM,YEAR,OLD/1+) 1-("y.old",ADD_).

Выявляются словосочетания, где вначале - число (слово с признаком NUM), затем слова "YEAR" и "OLD". Они преобразуются в число (1), к которому добавляется признак "y.old". Например, такое правило будет применимо к предложению "... mr. MILLS, 50 years old, ...".

Для терминов может быть задан допустимый контекст - слова или их признаки, стоящие слева и справа. Может быть также указан недопустимый контекст - слова или их признаки, которых не должно быть слева или справа. В результате удается выделять термины и словосочетания, значения которых зависят от контекста.

Для представления синонимов используются многоместные фрагменты:

SYNON(<результ.слово>,<исх.слово> ... <исх.слово>).

Например, SYNON(GRAPH,DIAGRAMM) - слово DIAGRAMM должно быть заменено на GRAPH. Эти же фрагменты служат для указания сокращений:

SYNON(CORP.,CORPORATION) SYNON(COMP.,COMPANY)

Синонимы, как правило, носят условный характер. Для них указывается допустимый или недопустимый контекст. Например, в приведенном выше случае недопустимы замены для слов - фамилий, кличек, названий улиц и др.

6. Контекстные правила

Блок синтактико-семантического анализа выполняет следующие функции:

- по признакам и контексту выделяет информационные или значимые объекты (ФИО людей, адреса, организации, номера машин и др.);
- для каждого выявленного значимого объекта находит в документе связанную информацию (для лиц это их год рождения, пол, адрес и др.).

Для этого используются "контекстные" правила.

Многие информационные объекты (адреса, номера машин, организации и др.) - это наборы слов, которые грамматически никак не согласованы. Их выделение может осуществляться по чисто формальным принципам. Например, адрес может рассматриваться как набор буквосочетаний 'P.O.', BOX, ST..., слов с большой буквы и чисел. Каждый такой набор может иметь свои границы и недопустимые компоненты. Например, в адресах не может быть ФИО, глаголов и т.д. Выделение таких наборов слов (описаний объектов) основано на использовании контекстных правил следующего вида:

CONTEXT(<слово1>,<слово2>,...,<словоN>) -> <результ. фрагмент>

где <слово1>,... это может быть - отдельное слово, признак, а также
И-ИЛИ графы. Для этих правил указывается, с какой позиции начинать применение, а также допустимый или недопустимый контекст. Далее, может быть указано, слово с какими признаками не должно стоять на той или другой позиции. Это обеспечивает дифференцированное применение правил. Все эти указания осуществляются с помощью фрагментов РСС.

Такие правила выделяют из текста группы слов (по их признакам), описывающих какой-либо объект, и заменяют их на одно слово, с которым связывается соответствующий фрагмент семантической сети, например, представляющий адрес.

Синтактико-семантический анализ предложений с выделением словосочетаний и анализом форм осуществляется на основе контекстных правил, которые применяются в определенной последовательности. Вначале выделяются объекты, затем их признаки, словосочетания, и наконец, глагольные формы. По мере применения таких правил строится семантическая сеть - содержательный портрет документа.

Применение каждого правила - это последовательность действий, основанных на анализе слов и их признаков. Например, рассмотрим, как применяется правило, выделяющее словосочетания с предлогом OF.

Правило содержит специальный фрагмент, который указывает, что применять это правило нужно с 2-ой позиции, т.е. искать слова OF. Другой фрагмент отделяет левую часть от правой (->). В правой части стоит фрагмент, который указывает, что слова на 1-й и 3-й позициях должны быть склеены в комбинацию слов, которое в дальнейшем будет рассматриваться как одно слово с признаком OBJ. Это правило осуществляет преобразования:

СЛОВО с признаком объект OBJ или англ. СЛОВО (с признаком ENGL) + OF + СЛОВО с признаком объект OBJ или англ. СЛОВО -> <комбинация слов>

Это пример наиболее простого правила. К таким правилам добавляются фрагменты, указывающие на контекст, на возможность каких-либо символов внутри и др. Специальные правила осуществляют идентификацию объектов, например, на основе местоимений или кратких описаний (по имени восстанавливается фамилия, если они где-нибудь упоминались вместе). И многое другое, что необходимо для работы с естественным языком.

Каждое контекстное правило - это семантическая сеть (РСС). Все лингвистические знания записываются в виде РСС. Над ними работают продукции языка DECL (программа), которые применяют эти правила и играют роль пустой лингвистической оболочки, поддерживающей язык записи лингвистических знаний - РСС. Как показывает опыт, такую оболочку можно настраивать на различные языки, т.е. строить различные лингвистические процессоры.

7. Применение правил

Контекстные правила применяются в строго определенной последовательности - каждое на своем уровне. Например, при обработке объявлений вначале выделяются информационные объекты - деньги с их количеством, даты, места событий др. Они сворачиваются и как бы представляют единое слово со своими признаками. Это необходимо, чтобы облегчить последующий анализ. Иначе слова, составляющие эти объекты, могут захватываться другими правилами и создавать шумы.

Далее начинается выделение лиц. Для этого вводится множество правил. Одни начинают свое применение с поиска имен, фамилий (MUSTBE), другие - с поиска года рождения, третьи - с инициалов. В результате минимизируются потери в случаях, когда блок морфологического анализа не дает необходимых признаков для каких-либо слов (что это имена или фамилии и т.д.). Затем анализируются словосочетания, и наконец, глагольные формы. По мере применения таких правил строится семантическая сеть - содержательный

портрет документа. Ниже приведен пример представления уровней, определяющих порядок применения правил.

```
{= Уровни =}

LEVEL(LEVEL_1,LEVEL_2,LEVEL_3,LEVEL_4,LEVEL_5)

LEVEL_1(MORF_ENG) {= Выявление частей речи англ. слов =}
LEVEL_1(MORF) {= Синонимы, термины =}

LEVEL_2(NNN~1,NNN~2) {= Выявление количества денег =}
LEVEL_2(TTT~1,TTT~2,TTT~3,TTT~4) {= Выделение дат =}
LEVEL_2(PPP~1,PPP~2) {= Выделение мест - PLACE_ =}
LEVEL_2(FFF~1,FFF~2,FFF~3,FFF~4) {= Выявление лиц =}
.....
LEVEL_3(GGG~1,GGG~2,GGG~3,GGG~4,GGG~5) {= Выявление словосочетаний =}

.....
```

В фигурных скобках даны комментарии.

В системе имеются контекстные правила, которые обеспечивают полный разбор предложений. Но в отличие от типовых грамматик параллельно обеспечивается выделение значимых (информационных) объектов, в том числе таких, в которых слова никак не согласованы между собой, например, адресов, машин с указанием их номеров и т.д.

8. Аналитические задачи.

Система ориентирована на анализ англоязычных документов с их автоматической формализацией и созданием пользовательской базы знаний. На этой основе возможно решение различных задач, в том числе, поиск информационных объектов, поиск похожих объектов, ответ на запросы в свободной форме, выявление связей объекта и др. [].

Представляется перспективным использование системы для поиска в сети ИНТЕРНЕТ. Для этого в экспериментальном варианте (на языке Perl) разработана приставка - для сбора информационных ресурсов. Её задача - найти в сети ИНТЕРНЕТ WEB-страницы, которые могут интересовать пользователя, и выделить из них части, могут содержать информационно-значимые объекты. Поиск осуществляется по ключевым словам, а выделение частей - по взвешенной сумме таких слов. На основе выделенных частей формируется пользовательская БЗ.

Другой класс задач - это формирование справок, аналитических отчетов на основе содержимого БЗ. Для настройки системы на пользовательские приложения (информационные объекты и формы их выдачи) была разработана специализированная оболочка - на языке обработки структур знаний DECL. В настоящее время такая настройка предполагает участие разработчика.

В качестве примера решалась задача сбора в сети ИНТЕРНЕТ сведений о продаже в США домов и земельных участков с аукциона в случае, когда семья не имеет возможности полностью возвратить кредит. Объявления о продажах размещают на специализированных сайтах. Объявления составлены в достаточно свободной форме и содержат много посторонней информации. В тоже время потенциальному покупателю нужны конкретные данные. Их выделение с формированием отчетов осуществлялось в соответствии с шаблоном: ШТАТ - ВРЕМЯ ПОСТУПЛЕНИЯ ЗАЯВКИ - СОБСТВЕННИК - ЦЕНА - МЕСТОРАСПОЛОЖЕНИЕ - ДАТА ПРОДАЖ - СВЯЗЬ - ОСОБЕННОСТИ УЧАСТКА.

Исходный материал - текстовые файлы, содержащие англоязычные тексты (они имеют вид - см. вышеприведенный пример части текста).

Результат:

Document 1. File ENG_1.TXX
State, County - GEORGIA COWETA COUNTY
Time of booking - 7 JANUARY 2002
Name of Owner - SHALMAR REESE DANIEL
Price - $9812
Address - 95 ALDER DRIVE NEWNAN GEORGIA 30263
Time of sale - 1 TUESDAY APRIL 2003
E-mail - WWW.FORECLOSUREHOTLINE.NET
Property - EXHIBIT УАФ ALL THIS TRACT OR PARCEL OF LAND CONSIST OF 0.409 ACRE LY AND BE LAND LOT 91 OF 5 LAND COUNTY GEORGIA SHOW BY PLAT OF SURVEY BY REGISTER LAND DATE SURVEYOR AND RECORD SURVEYOR PLAT BOOK 60 PAGE 79 OFFICE OF CLERK SUPERIOR COURT BE REFERENCE

Document 2. File ENG_2.TXX
State, County - GEORGIA COWETA COUNTY
Time of booking - 15 MAY 2002
Name of Owner - WILLIE AMEY DENE
Price - $57,600,
Address - 5 WEST PARK COURT NEWNAN GEORGIA 30263
Time of sale - 1 TUESDAY APRIL 2003
Property - ALL THIS TRACT OR PARCEL OF LAND LY AND BE CITY AND BE LOT 348 OF 3 SECTION UNIT 4 WESTGATE PARK SUBDIVISION RECORD SHOW PLAT PLAT BOOK 16 PAGE 111 RECORD BE REFERENCE WHAT PLAT HEREBY HAVE PARTICULAR DESCRIPTION

Document 3. File ENG_3.TXX
State, County - GEORGIA STATE COWETA COUNTY
Time of booking - 6 JANUARY 1998
Name of Owner - ANTHONY D. BERRY
Price - $12,450.53
Address - 79 SAVANNAH STREET NEWNAN GEORGIA 30263
Time of sale - 1 TUESDAY APRIL 2003
Property - ALL THIS TRACT OR PARCEL OF LAND LY AND BE CITY OF NEWNAN KNOW 79 SAVANNAH STREET ACCORD PRESENT SYSTEM NUMBER HOUSE HOUSE AND LOT IN SAY CITY AND DESCRIBE MORE FOLLOW BEGINN AT POINT ON SOUTHERLY SIDE OF SAVANNAH STREET WHAT SAY BE POINT NORTHWEST CORNER OF PROPERTY ....

Еще одно приложение связано с автоматическим составлением и пополнением статей электронной энциклопедии - на базе информации в ИНТЕРНЕТ. Соответствующие методики и программы разработаны Шарниным М.М., см. сайт WWW.KEYWEN.COM.

Литература

1. Кузнецов И.П. Методы обработки сводок с выделением особенностей фигурантов и происшествий. Труды международного семинара Диалог-1999 по компьютерной лингвистике и ее приложениям. Том 2. Тарусса 1999.

2. FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. AIC, SRI International. Menlo Park. California, 1996.

3. Кузнецов И.П. Семантические представления. М. Наука. 1986г. 290 с.
4. Kuznetsov Igor, Matskevich Andrey. System for Extracting Semantic Information from Natural Language Text. Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Том
2. Протвино, Наука, 2002.
5. Кузнецов И.П. Пузанов В.В., Шарнин М.М. Система обработки декларативных структур знаний ДЕКЛАР-2. Москва, ИПИАН, 1988 г.

СВЕДЕНИЯ ОБ АВТОРАХ:
Кузнецов Игорь Петрович, д.т.н., проф., ИПИ РАН, гл..н.с.,
Адрес: 105484 Москва, 16-парковая д.23 кв.220.
Тел. 461-30-97 (д), 135-43-80 (сл.)
E-mail:igor-kuz@mtu-net.ru

Мацкевич Андрей Георгиевич, ИПИ РАН, н.с.
Адрес: 123007 Москва, Хорошевское шоссе д.22 кв.53.
Тел. 941-14-69 (д), 135-43-80 (сл.)
E-mail:igor-kuz@mtu-net.ru

ABSTRACT

System Extracting Significant Information from Natural-Language Text in English.

Kuznetsov I.P., Matskevich A.G.
(Moscow, The Institute for Informatics Problems of the Russian Academy of Sciences) igor-kuz@mtu-net.ru

A system for extracting significant information from natural language text in English is considered. The system uses the methods which was developed for Russian language. In addition the programs of English morphological analysis was designed. Methods and programs of lexical and syntactic-semantic analysis of Russian language was adapted to English. Linguistic knowledge which provides the extraction of significant objects (persons, addresses and so on) from English texts was designed.

System oriented on automatic text collection, analysis and the user knowledge base forming. It can be used for various tasks: answer to question in free forms, searching the information objects, filling the data base, creating short reports about information objects and their links.