Часть III

ИНФОРМАЦИОННЫЕ РЕСУРСЫ
И ПОИСК ИНФОРМАЦИИ

УДК 681.3.01

ОРГАНИЗАЦИЯ СЕМАНТИКО-ОРИЕНТИРОВАННЫХ СИСТЕМ ПОИСКА И ОБРАБОТКИ ИНФОРМАЦИИ

И.П. Кузнецов

Рассматривается новый класс логико-информационных систем, использующих специальные лингвистические процессоры и технологию баз знаний (БЗ) для обработки потоков неформализованных документов с целью решения пользовательских задач. На первом этапе текст документа подвергается глубинной обработке с выявлением информационных объектов и связей. На основе последних формируются структуры знаний, которые образуют БЗ. На уровне БЗ осуществляется организация различных видов анализа и семантического поиска: похожих объектов, по связям и др. Рассматриваются основные компоненты подобных систем, названных семантико-ориентированными, и их конкретные приложения.

1. Актуальность проблемы создания и использования семантико-ориентированных систем обработки неформализованных документов

Лавинообразный рост потока документов, получаемых пользователями через различные информационные каналы (в том числе, из сети Интернет), требует новых решений для повышения эффективности поиска и анализа необходимой пользователям информации. Большая часть таких документов имеет вид текстов естественного языка (ЕЯ). Во многих случаях человек не в силах прочитать и осмыслить даже малую часть того, что ему предлагается. Существующие средства могут оказать помощь пользователям, но, как правило, требуют от них достаточно трудоёмкой работы.

В качестве примера рассмотрим две области, где имеют место существенные трудности такого рода.

Первая область — потоки документов в криминальной милиции. Следователь при решении конкретной задачи может найти полезную для себя информацию в различных источниках: сводках происшествий, справках по уголовным делам, обвинительных заключениях и др. В тоже время объёмы ежемесячной новой информации подобного рода исчисляются десятками и сотнями мегабайт. Никто не может всё это прочитать и удержать в голове.

Полнотекстовые базы данных не решают проблемы, так как при работе с текстами на ЕЯ дают много шумов (лишних документов) и потерь. Причина этого — особенности русского языка: наличие словоформ и свободный порядок слов. Одно и тоже понятие или событие можно описать множеством различных способов. Более того, слова запроса могут быть разбросаны по тексту документа и относиться к различным сущностям. И всё равно документ будет найден. Например, нужно найти Иванова Ивана, а в документе упоминаются Иванов Пётр и Петров Иван. Такой документ при поиске будет считаться адекватным.

Чтобы уменьшить процент шумов, используют различные методы: вводят критерии близости слов, обрезают окончания словоформ, вводят индексирование нормализованных слов и др. Но и это кардинально не решает проблемы.

Другой вариант — это использование реляционных БД. Но для этого требуются трудоёмкая работа специально обученных людей по формализации текстов на ЕЯ: выделение из текстового документа происшествия, лиц, адресов, дат и т.п., а также заполнение соответствующих таблиц БД. При больших потоках документов это сделать крайне трудно. В любом случае будут потери той информации, которая не учтена в рамках схем БД.

Описанная ситуация является типичной для многих других областей деятельности, связанной с анализом потоков информации в виде текстов на ЕЯ, поступающих через СМИ, ИНТЕРФАКС, из специальных источников.

Вторая область, это поиск в глобальной сети Интернет, где в настоящее время хранится огромное количество всевозможной информации. Подавляющее большинство документов — это тексты на ЕЯ. На данный момент в качестве помощи пользователю,

работающему в Интернет, предлагается класс поисковых машин, которые обеспечивают возможность контекстного поиска по ключевым словам запроса. Поисковая машина является универсальным инструментом и даёт много лишней информации, которую конечному пользователю приходится самостоятельно анализировать. Причиной этого является неспособность поисковой машины вылавливать то, что интересует пользователя.

Существенно не меняют картины и каталоги, с помощью которых можно найти конкретные Интернет-ресурсы в определённой предметной области. Каждый разработчик вынужден вручную добавлять свой ресурс в каталог. И хотя в ряде предметных областей существуют свои инструменты поиска, но и они дают много ненужного материала, что затрудняет работу с ними.

В тоже время большинство конкретных пользователей — это люди, которые интересуются конкретными вопросами. Например, следователю важны фигуранты, их места жительства, телефоны, криминальные события, даты и др. Специалиста по кадрам интересуют организации, где человек работал, кем он работал и когда это было. Другие люди вылавливают из СМИ информацию о странах, влиятельных лицах, катастрофах и др. Здесь важны и связи: места работы с занимаемой должностью, экстремальной ситуация с её временем и т. д.

Будем называть интересующую пользователя конкретную информацию — *информационными объектами*. Каждый пользователь (или класс пользователей) интересуется своими информационными объектами и связями между ними. Вся остальная информация является лишней и человек старается её просто не замечать. Отсюда часто используемая людьми методика чтения «по диагонали» или «с поиском ключевых слов».

Перспективное направление в области информатики — это обработка документов на ЕЯ, которая должна учитывать, прежде всего, интересы конечного пользователя. Отсюда следует необходимость построения нового класса информационных систем, использующих специальные *лингвистические процессоры* и технологию баз знаний (БЗ).

Лингвистические процессоры необходимы для глубинной обработки текстов с выявлением информационных объектов и связей. На основе последних формируются структуры знаний, которые образуют БЗ. На уровне БЗ становится возможным более полно учитывать потребности пользователя — за счёт организации различных видов поиска: конкретных объектов, похожих

объектов, по связям и др. Такие виды поиска относятся к «семантическим», так как осуществляются не на уровне слов или словоформ, а на уровне структур знаний из БЗ. Будем называть информационные системы подобного типа *семантико-ориентированными*.

Следует отметить ряд попыток их построения за рубежом [1]. В данной работе будет идти речь о проблемах построения, основных компонентах, структуре и перспективах использования семантико-ориентированных систем.

2. Структура семантико-ориентированных систем

На протяжении последних 15 лет в ИПИ РАН были разработаны различные классы семантико-ориентированных систем. Это комплексные системы ДИЕС, ИКС, «Аналитик». Рассмотрим их особенности на примере логико-аналитической системы «Аналитик» в приложении к задачам криминальной милиции [2]. Основные задачи этой системы: сбор всей поступающей информации (документов на ЕЯ), её автоматическая формализация и хранение, а также решение задач семантического поиска и анализа.

Система «Аналитик» ориентирована на автоматическую обработку документов в тех областях, где имеют место:

— большие потоки информации;

— неформализованный характер поступающей информации (это тексты на ЕЯ);

— высокая трудоёмкость формализации документов специально обученными людьми;

— необходимость исключить последствия недобросовестной работы людей при формализации документов.

Логико-аналитическая система «Аналитик» — это аппаратно-программный комплекс, автоматизирующий процесс ввода, формализации и анализа текстовых документов, их использование в задачах поиска и сложных видов обработки оперативной. Общая схема системы изображена на рис. 1.

Система содержит собственные базы данных и знаний, а также терминологический словарь. База данных (БД) системы «Аналитик» служит для хранения поступающих документов и структур знаний.

Система «Аналитик» обеспечивает автоматический ввод документов в БД с рабочих мест. По мере поступления документов автоматически выделяются ключевые слова и переводятся в ка-

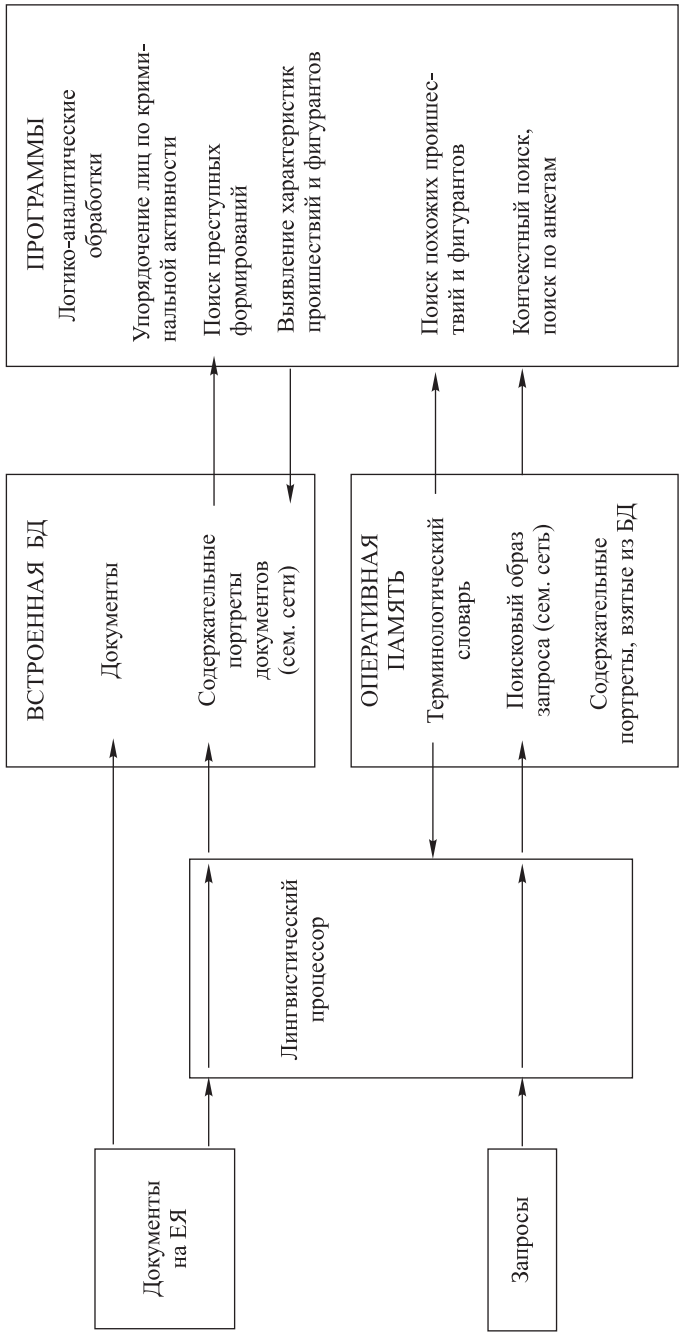

Рис. 1. Блок-схема семантико-ориентированной системы

нонической форму. На их основе строятся индексные файлы, обеспечивающие быстрый выбор документов.

База знаний системы «Аналитик» обеспечивает:

— хранение значимой информации и связей;

— эффективный поиск и анализ информации по связям.

Знания в БЗ представляются в виде структур, которые записываются в нотации семантических сетей (так называемых РСС), дополненных средствами представления событийных компонент и комплексных связей. В результате образуются так называемые *содержательные портреты*, смысл которых раскрывается ниже.

Содержательные портреты (как и документы) шифруются и помещаются в БД, ориентированную на большие потоки информации (сотни мегабайт) и обеспечивающую их быстрый выбор с дешифровкой — за счёт индексных файлов. Такие портреты подкачиваются в оперативную память по мере необходимости, образуя активную часть БЗ.

Для построения содержательных портретов (т. е. структур знаний) используется лингвистический процессор (см. рис. 1). Лингвистическая обработка включает в себя морфологический и синтактико-семантический анализ. За счёт первого обеспечивается нормализация элементов текста (приведение словоформ к одному виду, что очень важно для поиска), а за счёт второго — автоматическое выделение из него всей значимой информации: фигурантов, их примет, адресов, номеров их автомобилей, оружия и др.

Далее следует пост-лингвистическая обработка, заключающаяся в логическом анализе и выделении наиболее значимых характеристик документа: орудий преступления, способа его совершения, способа проникновения в здание или помещение и др. Осуществляется дополнение документа атрибутами — в соответствии с классификаторами, принятыми в криминальной милиции. Вся выделенная информация образует *содержательный портрет документа*, где представлены значимые элементы текста и их связи.

Такие портреты, представленные в нотации семантических сетей, образуют базу знаний, которая является основой для логико-аналитической обработки. За счёт выделения значимой информации, её дифференциации и использования связей удаётся повысить качество решения перечисленных ранее задач поиска и идентификации для оперативно-розыскной деятельности.

Логико-аналитическая обработка осуществляется на уровне структур знаний и ориентирована на логический анализ признаков, связей. Для этого используются соответствующие программы, что изображено в правой части рис. 1. Их задачи описываются ниже.

Терминологический словарь хранится в БЗ и определяет семантическое пространство терминов и признаков — с учётом их смысловой близости, синонимии и взаимного отрицания. За счёт этого расширяется пространство поиска, повышается точность и надёжность результатов, обеспечивается достаточная свобода использования слов и терминов в запросах и заданиях системе [4].

3. Семантические представления текстов на естественном языке

Структуры знаний должны с достаточной точностью обеспечивать представление семантической компоненты предложений ЕЯ. Отсюда следует выбор формализма представления. Вначале остановимся на общих соображениях.

Человек при решении логико-аналитических задач пользуется эвристическими методами, основанными на ассоциативных связях и собственном представлении о мире или о какой-либо предметной области, которое вырабатывается в процессе жизненного опыта. Такие представления во многих случаях носят обобщённый характер с привязкой к конкретным ситуациям. Например, понятия *разбойное нападение*, *злостное хулиганство*, *воровство* и др. могут быть расширены до многих конкретных сценариев, описываемых по-разному.

Итак, за словами человек видит реальные объекты и картины внешнего мира, присутствующие в них отношения (в широком понимании, где свойство — унарное отношение, а действие — k-местное). На этом уровне вырабатывается их сходство и различие, принадлежность к обобщённым сценариям. Поэтому, как правило, не важно, каким способом и в каких терминах выражается сценарий. Для этого зачастую может быть достаточно набора ключевых слов и понятий, характеризующих данный сценарий и позволяющий приблизительно восстановить картину.

Система не имеет представлений, подобных человеческим. В классическом треугольнике Фреге «внешний мир — представления о нём — язык» не хватает первого звена. Компьютер не видит внешнего мира. Поэтому форма системных представлений или знаний, на которые отображаются тексты и на уровне которых осуществляется обработка, закладывается разработчиком.

В соответствии с поставленными задачами важным этапом является выбор языка представления знаний, используемого для записи поисковых образов документов. Нужно учитывать, что слова (в силу высокого разнообразия способов выражения, см. ниже) не всегда характеризуют сценарии, которые как бы остаются в стороне. За счёт этого могут возникать значительные шумы и потери.

Наиболее адекватным средством представления и формализации сценариев в настоящее время являются *семантические сети*. Они могут быть различных типов. Наиболее перспективными, с точки зрения обработки текстов, являются семантические сети следующего вида.

Семантическая сеть состоит из множества вершин, представляющих объекты. Из вершин составляются элементарные фрагменты, каждый из которых представляет k-местное отношение. В этот фрагмент вводится две дополнительных вершины: одна соответствует отношению, а другая — всей совокупности упомянутых объектов с учётом их отношения. Эти вершины, как и любые другие вершины, могут стоять на местах объектов в других фрагментах, что обеспечивает высокие изобразительные возможности и гибкость: представление отношений между отношениями, между совокупностями связанных объектов и т. д.

Множество вершин делится на два подмножества: первое соответствует распознанным или определённым компонентам (именам, понятиям), а второе — неопределённым объектам, т. е. вопросительным словам, различного рода умолчаниям. Последние играют роль переменных.

Из таких фрагментов составляются сети, называемые *расширенными семантическими сетями* (РСС). Как показали исследования, подобные сети оказываются удобными для представления семантической компоненты различных языковых конструкций, в том числе, с отглагольными существительными и их формами, причастными оборотами, безглагольными конструкциями со связками типа «*это*, *есть*, *значит*» и др. [3].

4. Содержательные портреты документов

Содержательные портреты документов необходимы для обеспечения быстрого и качественного поиска информации по запросам, выраженным в достаточно произвольном виде.

Содержательный портрет — это расширенная семантическая сеть (РСС), которая представляет значимые объектов и их связи [6, 7]. Последние — это наборы сгруппированных признаков (слов в канонической форме).

Признаки группируются с учётом порядка их расположения в тексте, а также следующих факторов:

— какую информацию они представляют (приметы, адрес и др.);

— к какому значимому объекту относятся (лицу, машине, оружию и др.);

— в каком месте текста встретились и сколь близко расположены.

Отметим, что ориентация системы на определённые значимые объекты может легко меняться — за счёт изменения лингвистических знаний. Рассмотрим пример содержательного портрета документа, когда система ориентирована на выделение фигурантов, их примет, особенностей и совершаемых действий. Тогда при построении содержательного портрета из документа извлекается информация следующего вида:

— фигуранты, упоминавшиеся в документе (в том числе, неизвестные лица), каждому выявленному фигуранту присваивается свой код;

— ФИО каждого фигуранта (ФИО);

— приметы каждого фигуранта (ПРИМ_);

— национальность;

— адрес каждого фигуранта, где родился, прописан, где проживает (АДР.);

— место работы, должность каждого фигуранта;

— номера телефонов фигурантов;

— место и время происшествия;

— по какой статье проходит происшествие;

— марка и номер машины, её особенности;

— тип оружия, его номерной знак и другие особенности;

— соотнесённость с предложением (ПРЕДЛ_).

Содержательный портрет состоит из *фрагментов*. Это понятие шире, чем известное в логике понятие предикат. Каждый

фрагмент может иметь свой код, который может стоять на месте аргумента других фрагментов.

Рассмотрим содержательный портрет одного из документов. Пусть имеется следующий текстовый документ: «Обнаружен труп неизвестного мужчины с тремя огнестрельными ранениями в ногу, живот и область сердца. Давность трупа около 4 часов. Труп направлен в 11 морг. Его приметы: на вид 27–35 лет, рост 175–180 см., плотного телосложения, волосы чёрные, средней длины. Одет: пальто серое, коричневый пиджак, чёрная рубашка, синие брюки с манжетами».

Содержательный портрет этого документа имеет следующий вид: Содержательный портрет этого документа имеет следующий вид:

```
ДОК_(24,1-96.TXT,"Сводка;")
FIO(" "," "," ",1/1+) НЕИЗВЕСТНЫЙ(1-)
ПРИМ_(ВОЗРАСТ,27,35,РОСТ,175,180,КРЕПКИЙ,ТЕЛОСЛОЖЕНИЕ/2+)
ИМЕТЬ(1-,2-)
ПРИМ_(ПАЛЬТО,СЕРЫЙ,КОРИЧНЕВЫЙ,ПИДЖАК,ЧЁРНЫЙ,РУБАШКА,
      СИНИЙ,БРЮКИ,С/3+)
ИМЕТЬ(1-,3-)
КОЛИЧ_(3,ОГНЕСТРЕЛЬНЫЙ,РАНЕНИЕ/4+)
ОБНАРУЖЕНИЕ(ТРУП,1-,С,4-/5+)
ДАТА_("4:00"/6+)
ОРГ_(11,МОРГ/7+)
НАПРАВИТЬ(ТРУП,В,7-/8+)

ПРЕДЛ_(0,5-,В,НОГА,ЖИВОТ,И,ОБЛАСТЬ,СЕРДЦЕ)
ПРЕДЛ_(0,ДАВНОСТЬ,ТРУП,6-)
ПРЕДЛ_(0,8-)
ПРЕДЛ_(0,ОН,ПРИМЕТЫ,2-,"чёрные волосы",СРЕДНИЙ,ДЛИНА)
ПРЕДЛ_(0,ОДЕЖДА,3-,МАНЖЕТА)
```

Фрагмент `ДОК_(24,1-96.TXT,"Сводка;")` указывает на порядковый номер документа (24-й) и имя файла `1-96.TXT`, содержащего сводку с данным документом.

Фрагмент `FIO(" "," "," ",1/1+)` представляет неизвестное лицо — без ФИО. Если бы ФИО было известно, то на местах пробелов стояли бы фамилия, имя, отчество. Знак `1+` есть код фигуранта (код вводится через знак плюс). Знак `1-` это тот же самый код, но используемый повторно. С помощью таких кодов задаются отношения между объектами, а также места их расположения в предложениях.

Сказанное справедливо для знаков `2+`, `2-` и др.

Например, с помощью фрагмента ИМЕТЬ(1-,2-) представлены приметы ПРИМ_(ВОЗРАСТ,.../2+), которые относятся к неизвестному лицу. А с помощью ОБНАРУЖЕНИЕ(ТРУП,1-,С,4-/5+) представлено, что обнаружен труп неизвестного лица с тремя огнестрельными ранениями (код 4+, 4-).

Все объекты через их коды соотносятся к своим предложениям — ПРЕДЛ_(...), куда помещаются коды объектов (или действий над объектами), а также «лишние» слова, т.е. слова, не вошедшие в объекты. За счёт этого текст может быть восстановлен по содержательному портрету документа.

5. Уровни неоднозначностей, возникающих при формировании содержательных портретов неформализованных документов

Основная задача лингвистического процессора — отображение текстов в их содержательных портретах. При этом возникают существенные трудности, вызванные высоким разнообразием форм выражения, которые проявляется в текстах ЕЯ на различных уровнях.

Уровень словоформ — это слова в различных формах, которые значат одно и тоже, например, *борода*, *бороды*, *бороде* и т.д. Здесь необходимо учитывать словообразующие суффиксы, не изменяющие смысла слова и используемые для поддержания соответствующих языковых форм, например, *бородатый*, *бородатые* и т.д.

Для преодоления разнообразия на этом уровне используется морфологический анализ, который позволяет избавиться от различного написания слов, словоформ и использовать в поисковом образе документа каноническую форму слова (для существительных это именительный падеж, единственное число, для глаголов — инфинитив). В результате устраняются многие трудности, связанные с анализом и поиском.

Уровень понятий и терминов. При описании можно использовать термины различного уровня общности, например, *пистолет*, *огнестрельное оружие* и др. Такое разнообразие учитывается путём создания и использования в лингвистических знаниях синонимов, терминов, родовидовых или SUB-деревьев. При этом приходится учитывать случаи омонимии существительных и полисемии глаголов.

Здесь большую роль играет контекст. Например, *организация* — это может быть действие, а может быть и юридические лицо. Особое место занимает расшифровка сокращений путём анализа контекста. Например, Г. может означать ГОД, ГОРОД, ГОС. и др., в зависимости от контекста.

Уровень синтаксических или языковых форм. Одну и ту же мысль можно выразить по-разному: с помощью глагольных форм, отглагольных существительных, причастных оборотов и др. Например, факт «*ИПИРАН разрабатывает экспертные системы в области криминалистики*» может быть выражен следующим образом:

— ИПИРАНовская разработка экспертных систем в области криминалистики;

— Разработчиком экспертных систем в области криминалистики является ИПИРАН;

— ИПИРАН, который разрабатывает экспертные криминалистические системы.

Для преодоления разнообразия на уровне языковых форм и выявления фактов используются синтаксический и семантический анализ. Синтаксический анализ необходим для выделения словосочетаний, связанных групп слов, актантов глагольных форм. Он позволяет использовать в поисковом образе более сложные языковые конструкции — группы существительного, глаголов, генитивные цепочки слов.

Семантический анализ необходим для выделения объектов, о которых идёт речь в документе, их ролевых функций и связей между объектами. Это уже факты, конкретные сведения. Система как бы отвлекается от слов, имён, терминов: каждый из них может встречаться многократно и обозначать различные объекты. На этой основе строится содержательный портрет документа, для формализации которого используются семантические сети. Последние образуют базу знаний, которая обеспечивает фактографический поиск и сложные виды аналитической обработки — на основе связей.

Уровень описания происшествия или сценария. Один и тот же сценарий описывается различными людьми совершенно по-разному с акцентацией на различные стороны с использованием слов и глагольных форм, описывающих различные компоненты и отношения между ними, умолчанием очевидных фактов.

Если трудности, возникающие на первых трёх уровнях, в какой-то степени преодолимы, то четвёртый уровень — чисто эвристический, который учесть практически невозможно.

6. Семантико-ориентированный лингвистический процессор

Задачи семантико-ориентированного лингвистического процессора (ЛП) заключаются в следующем:

— преобразование слов в каноническую форму (морфологический анализ);

— выделение из документа информационных объектов и связей с преобразованием в структуру, удобную для последующей обработки.

При этом используется **семантико-ориентированный подход**. Его особенность в том, что система как бы старается быстро отвлечься от чисто языковых явлений, перенося обработку на семантический уровень. Иначе говоря, используется подход, характерный для человека, который за счёт смыслового анализа хорошо понимает неправильно построенные предложения, зачастую состоящие из отдельных ключевых слов, например, «*Моя иметь квартира*».

Путём использования мощных средств анализа и специального терминологического словаря в системе устраняются многие негативные явления, вызванные разнообразием языковых форм и терминов, а также неоднозначностями, возникающими на различных уровнях анализа (см. п. 5).

Для обработки используются **продукционные средства** — наборы правил: «ЕСЛИ..., ТО...», специально ориентированные на работу с расширенными семантическими сетями (РСС). Левая или условная часть каждого правила определяет вид анализа, а правая — действия при выполнении условия. В левой части могут стоять любые наборы фрагментов с переменными (они принимают значения в процессе применения правил), а в правой — добавляемые фрагменты и встроенные операторы.

Эти правила достаточно независимы. Их можно легко менять. Таким путём обеспечивается гибкость, возможность построения алгоритмов с высокой глубиной анализа и их быстрой подстройкой под конкретного пользователя [5].

Таким образом, вся содержательная обработка осуществляется на уровне семантических сетей с помощью правил: «ЕСЛИ..., ТО...».

Семантико-ориентированный ЛП выявляет из документов, по возможности, все объекты и связи между ними с автоматическим построением структур знаний (в виде семантических сетей) и их использованием для фактографического поиска и логико-аналитической обработки.

По каждой глагольной или же какой-либо другой языковой форме строится фрагмент, представляющий соответствующее отношение или действие с указанием роли выявленных объектов. **Квантование предложений** (их разбиение на связанные группы слов) идёт с использованием **семантических критериев** — по их участию в отношениях или действиях.

Особенности семантико-ориентированного ЛП состоят в следующем:

— поддержка модели языка с учётом семантических характеристик слов и словообразующих компонент;

— морфологический анализ слов с учётом приставок, словообразующих суффиксов и отглагольных форм;

— синтаксический и семантический анализ текстов, выделение объектов, их признаков и связей с автоматическим формированием структур знаний — семантических сетей;

— наличие предметно-ориентированных словарей и родовидовых деревьев, необходимых для семантического анализа текстов;

— анализ анафорических ссылок (местоимений) с идентификацией соответствующих объектов;

— выделение признаков, связей, относящихся к описываемому значимому объекту, сбор сведений об объекте;

— восстановление информации, данной по умолчанию (например, фраза «*деревянный дом*» означает, что *дом сделан из дерева*);

— поиск для каждого значимого объекта близких ему объектов (критерии близости — наличие одинаковых свойств, участие в аналогичных отношениях или действиях в определённых ролях).

При обработке дополнительно порождается, выявляется и учитывается аналитическая информация, характеризующая документ и выделенные информационные объекты. Это осуществ-

ляется путём использования терминологического словаря, представленного тоже в виде семантической сети. С этой целью вводится этап **пост-лингвистической обработки**.

7. Семантические фильтры лингвистического процессора

Семантические фильтры являются составной компонентой лингвистического процессора. Они обеспечивают (на этапе постлингвистической обработки за счёт использования терминологического словаря) содержательный анализ информации документа с пополнением его содержательного портрета значимыми фактами и характеристиками. Применительно к предыдущему примеру это — автоматическое выявление из текстов описания атрибутов фигуранта, его словесного портрета, формирование по классификатору особенностей происшествия.

Семантические фильтры основаны на идеологии *фреймов*. Каждый семантический фильтр — это обобщённая форма, в которой имеются уточняемые компоненты (в терминологии фреймов это «слоты», которые заполняются конкретными объектами или их описаниями). Для уточнения используются родовидовые деревья, называемые *SUB-деревьями*.

Парадигма семантических фильтров в области искусственного интеллекта достаточно хорошо известна под названием «демоны за круглым столом». В данном случае эта парадигма будет использоваться для расширения множества анализируемых признаков, целенаправленного вовлечения в качестве признаков значимой информации, а также для решения других важных задач текстовой обработки: автоматического анализа текстовых документов и заполнения соответствующих полей информационных карточек для существующих БД.

Важным элементом семантических фильтров являются SUB-деревья, а также компоненты, задающие семантические пространства. SUB-деревья, которые состоят из классов, подклассов и значимых объектов, связанных отношением «род–вид» (в ряде случаев в рамках SUB-дерева удобно представлять связь типа или «часть–целое»). Такое SUB-дерево включает в себя основные разделы классификатора: преступные действия, оружие, должностные лица, характер связи и др.

Каждый такой раздел расшифровывается. Например:


```
ОРУЖИЕ:
  - Взрывчатое вещество
  - КАСТЕТ,
  - ПИСТОЛЕТ,
  - ВАЛЬТЕР,
  - ПИСТОЛЕТ ПМ,
  - ПИСТОЛЕТ ТТ,
    ......
```

Значимые глаголы (а также другие языковые средства, выражающие интересующие пользователя оценки или явления) делятся на группы или семантические пространства, в каждый из которых помещаются глаголы с похожими действиями или одинаковыми результатами.

Пример 7.1. «*Отсутствовать*, *без*, *остаться без*, *разогнать*, *распустить*, *кончился срок*» — у всех этих действий одинаковый результат. Они образуют группу с основным словом ⟨отсутствовать⟩.

На этой основе строятся семантические фильтры. Типичный пример семантического фильтра выглядит следующим образом:

⟨страна или её часть⟩ ⟨отсутствовать> ⟨орган власти>.

Таким фильтром охватываются различные способы выражения близких по смыслу компонент текста: «*Чечня без парламента*; *Разогнанный парламент Чечни*; *Чечня*, у которой распущен парламент; У парламента Чечни закончился срок и др.».

Вместо слов *Чечня* и *парламент* могут стоять другие слова, допустимые в семантическом фильтре.

Отметим некоторые наиболее важные моменты, связанные с построением семантических фильтров.

Во-первых, при построении SUB-деревьев не требуется перечисления всех элементов каждого класса, например, всех видов *оружия*, *органов власти* и др. Важно, чтоб были наиболее типовые элементы, часто встречающиеся в текстах происшествий или описаниях фигурантов, а значит, понятные простому читателю. Достаточно, чтоб SUB-деревья покрывали 90–95 % случаев (так как фильтры используются в качестве дополнительных признаков). Подобные SUB-деревья строятся достаточно быстро и просто.

Во-вторых, в ряде случаев классы можно пополнять автоматически, пользуясь контекстом. Например: «*автомашина* ⟨...⟩, *город* ⟨...⟩, *село* ⟨...⟩, *президент* ⟨...⟩» и др. Любое слово,

стоящее на месте многоточия и начинающиеся с большой буквы, может быть автоматически отнесено к соответствующему классу. Более того, новые города, посёлки, районы также можно автоматически относить к республикам или странам, так как последние обычно упоминаются чуть ранее.

В-третьих, объекты, выражаемые многосложными формами, можно вводить через семантические фильтры. Например, фильтры:

⟨особые приметы⟩ ::= ⟨цвет⟩ ⟨особые приметы⟩
⟨одежда⟩ ::= ⟨цвет⟩ ⟨одежда⟩
⟨одежда⟩ ::= ⟨материал⟩ ⟨одежда⟩

позволяют распознавать словосочетания типа «*рыжая борода*, *синие джинсы*, *кожаная куртка*» и относить их к особым приметам или *одежде*.

Каждый такой фильтр строится для того, чтобы выделять интересный материал или явление. Например, в настоящее время построена система фильтров для выделения признаков фигурантов, характеризующих их словесные портреты. Они могут быть также использованы для заполнения соответствующих информационных карточек фигурантов по текстам их описания, встречающихся в различных документах.

8. Терминологический словарь

Терминологический словарь служит, во-первых, для выявления особенностей документа и его значимых объектов (при использовании в семантических фильтрах), и во-вторых, для расширения пространства поиска и формирования объяснительной компоненты (при поиске и логическом выводе). Он также обеспечивает представление типовых классификаторов, служащих для различения особенностей происшествий и фигурантов. Словарь содержит ключевые понятия (классы), связи между ними, представленные в нотации семантических сетей.

Терминологический словарь представляется в виде структур знаний — фрагментов семантической сети. Он содержит следующие виды связей:

— род–вид, класс–подкласс (для представления таких связей в семантической сети используются фрагменты типа SUB);

— безусловные синонимы (используются фрагменты SYNON);

— условные синонимы, т. е. слова совпадают по смыслу при определённом контексте (фрагменты SYNON);

— антонимы, т. е. противоположные по смыслу (фрагменты OR_OR);

— взаимоисключающие (фрагменты OR_DR);

— близкие по смыслу, т. е. из одного вытекает другое (фрагменты NEAR);

— представляющие семантические фильтры (используются фрагменты WORD).

— образующие словосочетания (фрагменты WORD).

Рассмотрим фрагмент семантического словаря на следующем примере.

```
   SUB(ГЛАЗА,СЛЕПОЙ)
SYNON(СЛЕПОЙ,ОСЛЕПНУТЬ,СЛЕПНУТЬ,ПОДСЛЕПОВАТЫЙ)
   SUB(ГЛАЗА,"глаза нормальные")
SYNON("глаза нормальные",НОРМАЛЬНЫЙ,ОБЫЧНЫЙ/1+) 1-(ГЛАЗА)
   SUB(ГЛАЗА,"плохо видит")
WORD("плохо видит",ПЛОХОЙ,ВИДЕТЬ)
OR_OR("глаза нормальные",СЛЕПОЙ,КОСОГЛАЗЫЙ,"плохо видит")
NEAR(СЛЕПОЙ,"плохо видит")

WORD("Особые приметы",ЦВЕТ,"Особые приметы")
          SUB(ЦВЕТ,ЧЕРНЫЙ) NEAR(ЧЕРНЫЙ,ТЕМНЫЙ)
          SUB(ЦВЕТ,ТЕМНЫЙ)
          SUB(ЦВЕТ,СЕРЫЙ)
```

Фрагмент

```
SYNON(СЛЕПОЙ,ОСЛЕПНУТЬ,СЛЕПНУТЬ,ПОДСЛЕПОВАТЫЙ)
```

означает, что слова-признаки являются синонимами. Система использует такие фрагменты, чтоб приводить слова-признаки в содержательных портретах входных документах к одному виду. Имеется в виду слово, которое стоит во фрагменте на первом месте.

Фактически, таким способом устраняются недостатки блока морфологического анализа.

Фрагмент

```
SYNON("глаза нормальные",НОРМАЛЬНЫЙ,ОБЫЧНЫЙ/1+) 1-(ГЛАЗА)
```

— это условный синоним. За счёт этого фрагмента в содержательном портрете документа слово *нормальный* или *обычный* будет заменено на признак «глаза нормальные» только, если рядом (в пределах 2–3-х позиций) стоит слово *глаза*. Условные

синонимы необходимы, так как слова типа *нормальный*, *обычный*, могут относиться к чему угодно.

Фрагмент типа `NEAR(...)` указывает на близость признаков и используется для расширения пространства поиска, а также при выявлении аналитических признаков. Система вместо слов текста пробует подставлять близкие слова и пробует таким способом искать адекватные документы или выявлять конкретные данные и факты.

Фрагменты типа `OR_OR(...)` означают или то, или другое, или третье. Они используются в различных видах аналитической обработки для выявления несоответствий, противоречий.

Фрагмент `SUB(ГЛАЗА,СЛЕПОЙ)` представляет отношение «род–вид». Он означает, что глаза могут быть слепыми. Такие фрагменты служат для отнесения информации к определённому классу. Они образуют SUB-дерево, представляющее ветви классификатора.

Концепция семантических фильтров реализуется с помощью фрагмента вида

```
WORD("Особые приметы",ЦВЕТ,"Особые приметы")
```

где `ЦВЕТ` и `"Особые приметы"` имеют пояснения. Признак `"Особые приметы"` будет сформирован при наличии рядом стоящих слов, относящихся к классам `ЦВЕТ` и `"Особые приметы"`. Такие слова могут стоять в любом порядке на расстоянии в пределах 2–3-х позиций, что позволяет учесть разнообразные языковые формы с различными словами.

За счёт последнего фрагмента и ветвей SUB-дерева примера 2 словосочетания типа *чёрная маска*, *рыжая борода* также будут отнесены к классу `"Особые приметы"`.

Отметим два важных момента. Во-первых, фрагменты типа `SUB(...)`, `OR_OR(...)` и `NEAR(...)` играют важную роль для расширения пространства поиска. На базе имеющихся слов-признаков запроса порождаются так называемые *вторичные признаки*:

— близкие по смыслу термины (на основе фрагментов `NEAR`);

— поясняющие термины (на основе фрагментов `SUB`);

— наличия противоречивых признаков (на основе фрагментов `OR_OR`).

В результате в поиск вовлекается значительно большее число признаков.

Во-вторых, фрагменты типа `SUB(...)` и `WORD(...)` фактически представляют собой обобщённые знания, которые позволяют выявлять качественные характеристики в соответствии с родовидовым деревом. В настоящее время подобные знания строятся человеком. В перспективе фрагменты типа `WORD(...)` предполагается строить на основе *обучающей выборки*.

Ниже в качестве примера будут рассмотрены две семантико-ориентированные системы, доведённые до уровня реальных приложений.

9. Система автоматической формализации текстов с выдачей результатов на естественном языке

Система LINGVO-MASTER, обеспечивающая автоматическую формализацию различного рода справок и сообщений (автобиографических данных, заявок на работу, резюме, сообщений СМИ), представляющих собой тексты естественного языка. При этом используется методика, состоящая из четырёх этапов.

На первом этапе вызывается блок морфологического анализа, который преобразует текст в семантическую сеть, представляющую поверхностную структуру текста. В этой сети все слова преобразованы в каноническую (нормальную) форму. В ней представлен порядок расположения слов и других знаков, а также начало и конец каждого предложения. Для каждого слова указаны его морфологические характеристики (часть речи, падеж и др.) и чисто формальные признаки: слово с большой буквы, большими буквами, с точкой на конце и др.

На втором этапе работает *прямой лингвистический процессор* (ЛП), который управляются с помощью лингвистических знаний (ЛЗ). Прямой ЛП — это оболочка, которая становится процессором после её заполнения ЛЗ. ЛЗ — это также структуры знаний, т.е. РСС. Прямой ЛП преобразует тексты в структуры знаний, представляющие собой содержательные портреты этих текстов.

На третьем этапе полученная структура дополняется новыми данными. Например, при формализации анкетных данных человека устанавливается (по соответствующему классификатору) профессиональная область, степень его образованности, знание языков и др. Это осуществляется с помощью экспертных си-

стем (ЭС), которые представляют собой оболочку с экспертными знаниями. В соответствии с ними ЭС анализируют фрагменты содержательного портрета для выбора того или иного пункта классификатора, например, «иметь высшее образование» или «неполное высшее».

На четвёртом этапе работает *обратный лингвистический процессор* (ЛП), который преобразует структуры знаний в компоненты ЕЯ и отображает их на поля анкеты или сайта. Обратный ЛП — это также оболочка, которая управляется своими ЛЗ. ЛЗ обратного ЛП определяют вид формируемой анкеты и связь её полей с компонентами содержательного портрета документа.

Рассмотрим это на следующем примере:

РЕЗЮМЕ

Фалетова Наталья Ивановна

+7(910)412-88-57 (моб.),592-57-61 (дом.)
faletov@mail.ru
30 лет.
Образование
Московский Государственный Строительный Университет
Специальность: Инженер-строитель-технолог

Опыт работы:
2003-2005 ООО "Бизнесцентр "Алмазный"
Начальник отдела аренды нежилых помещений
Руководство отделом (5 человек), ведение переговоров

05.2002 - 11.2003г. ЗАО "Сооружение"
Начальник отдела договорной работы (нежилые помещения)
Руководство отделом, ведение переговоров

Заработная плата от 900 у.е.

В результате обработки данного РЕЗЮМЕ формируется следующая форма:

Язык резюме:–> Русский
Фамилия:–> *Фалетова*
Имя:–> *Наталья*
Отчество:–> *Ивановна*
Пол:–> Женский
Дата рождения:–>
Возраст:–> 30
Образование:–> ## 1 — (Высшее)

Желаемая должность:–>
Профессиональная область:–>
Специализация:–>
Опыт работы:–> ## 1–(1–3 года)
Ожидаемый уровень месячного дохода:–> *От 900 евро*
Регион:–>
Возможность переезда:–>
E-mail:–> faletov@mail.ru
Почтовый адрес:–>
Домашний телефон:–>592-57-61
Мобильный телефон:–> +7(910)412-88-57
Рабочий телефон:–>
Телефон:–>
Личная интернет-страница:–>
УЧЁБА:–>
Название:–> *Московский Государственный Строительный Университет*
Факультет (специальность):–> *Инженер строитель технолог*
Диплом (степень):–>
Начало учёбы:–>
Окончание учёбы:–>
ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:–>
Начало работы:–> 2003
Окончание работы:–> 2005
Название организации:–> *ООО Бизнесцентр Алмазный*
Занимаемая должность:–> *Руководство отделом. Начальник отдела аренды нежилых помещений*
Обязанность, функции, достижения:–> *Руководство отделом (5 человек), ведение переговоров*
ПРОФЕССИОНАЛЬНЫЙ ОПЫТ:–>
Начало работы:–> Май 2002
Окончание работы:–> Ноябрь 2003
Название организации:–> *ЗАО Сооружение*
Занимаемая должность:–> *Руководство отделом. Начальник отдела договорной работы (нежилые помещения)*
Обязанность, функции, достижения:–>
КУРСЫ (обучение):–>
ЯЗЫКИ:–>

Другое приложение системы LINGVO-MASTER это анализ текстов, выявление Объектов и заполнение ими полей БД.

10. Логико-аналитическая система «Криминал»

Система «Криминал» разработана на базе системы «Аналитик» и предназначена для областей, где имеют место потоки текстовой информации: сводки происшествий, СМИ, сообщения о новом оборудовании, дефектах, катастрофах, организациях, ценах и др. Автоматизирует процессы, связанные с созданием баз данных и знаний, семантическим поиском, составлением отчётов и др.

Типовые задачи пользователя.

— Поиск в этом потоке интересующих его информационных объектов. Это могут быть лица, организации, те или иные виды оборудования, их особенности (дефекты), события определённого типа (криминальные действия, сбои оборудования, изменение цен на товары,...), их место, время и др. Каждая область приложений характеризуется своими объектами.

— Выявление связей объектов. Например, как интересующие его лица связаны с организациями, кто принимал участие в событиях, когда они имели место (точные даты) и др.

— Составление на этой основе отчётов, протоколов.

Ядро системы «Криминал» содержит следующие основные компоненты:

Уникальный лингвистический процессор, который обеспечивает:

— автоматическую формализацию текстовой информации на русском языке с выявлением лиц, организаций, промышленных изделий, событий, дат и др., их связей и создание на этой основе собственной базы знаний (БЗ);

— автоматическое построение каталогов информационных объектов.

— ввод данных в БЗ через анкеты;

— автоматическое заполнение информационными объектами тематических полей Базы Данных (в автономном режиме).

Лингвистический процессор содержит программное ядро, работа которого определяется лингвистических знаний. Настройка на предметную область — только за счёт лингвистических знаний.

Сервисные программы, которые на основе информации в базе знаний обеспечивают решение логико-аналитических задач на основе информации в БЗ:

— ответ на запросы в свободной форме (на ЕЯ);

— поиск связей между объектами;

— выявление и ранжирование объектов по качественным критериям, заданным пользователем (криминальная активность и др.);

— различные виды поиска информационных объектов по базе, в том числе нечёткий поиск, поиск похожих событий и др.

— режим гипертекста — для перебора мест вхождения выбранного информационного объекта (по –>);

— автоматическое построение графических схем, отчётов, диаграмм, отражающих особенности интересующих пользователя объектов.

Формируемые системой каталоги позволяют быстро находить нужный объект и быстро решать на это основе перечисленные задачи.

В системе имеются средства, с помощью которых обеспечиваются различные настройки на виды поиска, способы оценки и решение различных задач.

11. Заключение

Семантико-ориентированные системы обработки неформализованной информации, представленной в виде текстов на естественном языке, — это перспективное направление в области информатики с широким кругом приложений. Помимо упомянутых ранее задач, такие системы могут быть использованы для дифференцированного извлечения информации из сети Интернет. В этом случае по запросу пользователей, выраженному в свободной форме, обеспечивается дифференцированный поиск в сети Интернет необходимой информации, выделение из неё интересующих пользователя компонент, их содержательный анализ с выдачей пользователю результатов в наиболее удобном и сжатом виде, например, в виде рефератов или форм с заполняемыми полями.

Другие возможные приложения: анализ потока сообщений, выявление полезной информации и её накопление в базе знаний с последующим использованием для постоянного информирования пользователя в его предметной области.

В данном варианте поддерживаются различные виды поиска, в том числе нечёткого, а также запросы в свободной форме на естественном языке.

Обеспечивается также и решение аналитических задач: поиск по связям, ранжирование объектов по степени их активности, анализ изменений регулярных событий во времени и др.

Список литературы

1. FASTUS:a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. AIC, SRI International. Menlo Park. California, 1996.
2. *Кузнецов И.П.* Методы обработки сводок с выделением особенностей фигурантов и происшествий. Труды международного семинара Диалог-1999 по компьютерной лингвистике и её приложениям. Т. 2.Тарусса, 1999.
3. *Кузнецов И.П.* Семантические представления. — М.: Наука, 1986. — 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 её приложениям. Т. 2. Протвино, Наука, 2002.
5. *Кузнецов И.П.*, *Пузанов В.В.*, *Шарнин М.М.* Система обработки декларативных структур знаний ДЕКЛАР-2. — М.: ИПИАН, 1988.
6. *Кузнецов И.П.* Особенности обработки текстов естественного языка на основе технологии баз знаний // Сб.н. тр. ИПИ РАН. Вып. 13. — М.:Наука,2003. — С. 241–250.
7. Igor Kuznetsov, Elena Kozerenko. The system for extracting semantic information from natural language texts // Proceeding of International Conference on Machine Learning, MLMTA-03. Las Vegas US, 23–26 June 2003. P. 75–80.