В. Ф. Хорошевский, Пространства знаний в сети Интернет и Semantic Web. Часть 3, *Искусственный интеллект и принятие решений*, 2012, выпуск 1, 3–38

В.Ф. Хорошевский

Пространства знаний в сети Интернет и Semantic Web. Часть 3

Аннотация. Работа является завершением цикла статей, где представлено направление Семантический Веб. В ней дан аналитический обзор решений, продуктов и систем, разрабатываемых в России. Изложение структурировано по трем направлениям – инструментальные средства представления знаний, методы извлечения информации из текстов, а также пространства знаний для прикладных интеллектуальных систем, функционирующих в среде Интернет.

Ключевые слова: Семантический Веб, пространство знаний, обработка естественного языка, извлечение информации из текстов, инструментальные средства, продукты и сервисы, аналитический обзор.

Введение

Настоящая работа является завершением цикла статей автора [1, 2], в которых было представлено направление Семантический Веб, дан общий обзор методов и средств извлечения информации из текстов на примере анализа решений и систем, разрабатываемых зарубежными коллективами и организациями, а также представлена общая картина исследований по этой тематике в России и некоторых странах СНГ. В данной статье более подробно обсуждаются решения, продукты и системы, разрабатываемые в России. При этом автор опирается на результаты анализа активностей исследовательских коллективов и коммерческих компаний, полученные в работах [2, 3], материалы, любезно предоставленные коллегами, а также информацию из Интернет.

Изложение организовано следующим образом: для относительной самодостаточности настоящей статьи, сначала кратко резюмируются результаты работ [2, 3], а затем регулярным образом обсуждаются наиболее интересные, на наш взгляд, аспекты российских исследований и разработок в области извлечения информации из ЕЯ-текстов, а также прикладные системы и сервисы для Семантического Веба.

Как уже отмечалось в [2], спектр российских организаций и коллективов, работающих в области обработки ЕЯ, имеет специфический качественный и количественный состав. Во-первых, исследования и разработки в данной области ведутся в большом числе небольших и малых исследовательских коллективов (около 100 проектов, коллективов, организаций, причем в каждом из них участвует 3-5, редко 10 человек) и в очень ограниченном числе коммерческих организаций. Во-вторых, исследования, в большинстве своем, имеют теоретический характер. И только единичные исследовательские коллективы доводят их результаты до реально действующих систем, а коммерческие организации, позиционирующиеся в данном секторе, в большинстве своем, сосредоточены лишь на решениях, которые можно быстро вывести на рынок, даже за счет снижения наукоемкости и качества соответствующих продуктов. И, наконец, российские исследования и разработки в данной области характеризуются недостаточностью достоверной информации о коллективах, здесь работающих, о концепциях, «исповедуемых» разными коллективами, методах реализации разрабатываемых систем и реально полученных результатах. Поэтому следующий ниже аналитический

обзор базируется, в основном, на обсуждении работ «скрытых» коллективов, выявленных в результате автоматической обработки публикаций российских авторов в трудах ведущих конференций по данной тематике. К сожалению, этот анализ может оказаться неполным и/или фрагментарным, но, даже понимая это, автор считает полезным представить его заинтересованному читателю.

Как показано в предшествующих работах данного цикла, в области разработки методов и систем извлечения информации из текстов на естественных языках в интересах создания пространств знаний, а также использования сформированных БЗ в рамках прикладных информационно-аналитических систем и/или для обслуживания семантических Интернет-сервисов выделяются кластеры, представленные в первых пяти позициях Табл.1, к которым добавлены кластер инструментальных средств автоматической обработки ЕЯ-текстов (SDK) и кластер семантических продуктов и продуктов для Семантического Веба (SP&SW).

Заметим, что в кластер SP&SW были выделены те организации, которые уже прошли стадию демонстрационных прототипов и выходят, как минимум, на бэта-тестирование своих продуктов, а также продукты и системы организаций, которые в информационном поле научных конференций практически не представлены.

И, наконец, последнее замечание перед тем, как перейтии к более детальному анализу ситуации. В настоящей работе основное внимание уделяется прикладным исследованиям по извлечению информации из ЕЯ-текстов в контексте формирования пространств знаний и Семантического Веба и соответствующему инструментарию. Поэтому, а также в силу ограничений на объем статьи, из дальнейшего рассмотрения полностью исключены кластеры IR, CC, MPKS и KBS, а кластер OE, в силу его значимости для приложений Семантического Веба, предполагается обсудить в отдельной статье.

1. Российские исследования и разработки

1.1. Карта леса

Общая структура исследований и разработок в области извлечения информации из текстов на естественных языках, как она представляется на основании анализа публикаций в этой области, может быть представлена схемой, показанной на Рис.1 .

Учитывая специфику российских исследований и разработок в данной области, дальнейшее обсуждение организовано следующим образом: сначала рассматриваются геоландшафты российских исследовательских коллективов и организаций, работающих в области извлечения информации из ЕЯ-текстов, затем, более подробно, обсуждается инструментальное направление и собственно системы извлечения информации, а в завершающей части статьи представлены некоторые продукты и сервисы для Семантического Веба.

Ретроспективный анализ статей по тематике обработки ЕЯ за 2000-2010 г.г., опубликованных в трудах основных российских конференциях (ДИАЛОГ, РАИИ и РОМИП), показывает [2, 3], что лишь около 20% участников этих конференций позиционируются в домене исследований и разработок, связанных с прикладными аспектами обработки ЕЯ-текстов. Исключение составляют конференции РОМИП, где этот показатель близок к 100%, так как это, по существу, российский TREC [4]. Определенный интерес в этом контексте, на наш взгляд, представляет организационно-террито-

Табл.1

Кластеры	Организации (по алфавиту)
Информационный поиск (IR)	АНО ЦИИ, Галактика, ИСА РАН, МГУ, Яндекс, RCO, Mail.Ru
Кластеризация и классификация (CC)	АНО ЦИИ, ГУ ВШЭ, ИСА РАН, РГГУ, Яндекс
Онтологический инжиниринг (OE)	Авикомп Сервисез, АНО ЦИИ, ИСИ СО РАН, МГУ, РосНИИ ИИ, СПбПГУ
Извлечение информации из текстов (IE)	Авикомп Сервисез, ИПИ РАН, МГУ, RCO
Методологические проблемы (MPKS)	СПбГУ, РГГУ
Системы, основанные на знаниях (KBS)	ИПИ РАН, ИСА РАН, ИСИ СО РАН, КГУ, МГУ, RCO
Инструментальные средства (SDK)	Авикомп Сервисез, ВМиК МГУ, ИПИ РАН, RCO
Семантические продукты и Семантический Веб (SP&SW)	Авикомп Сервисез, АНО ЦИИ, Галактика-Zoom, Медиалогия, Синергетические системы, Сайтек, i-Teco, RCO

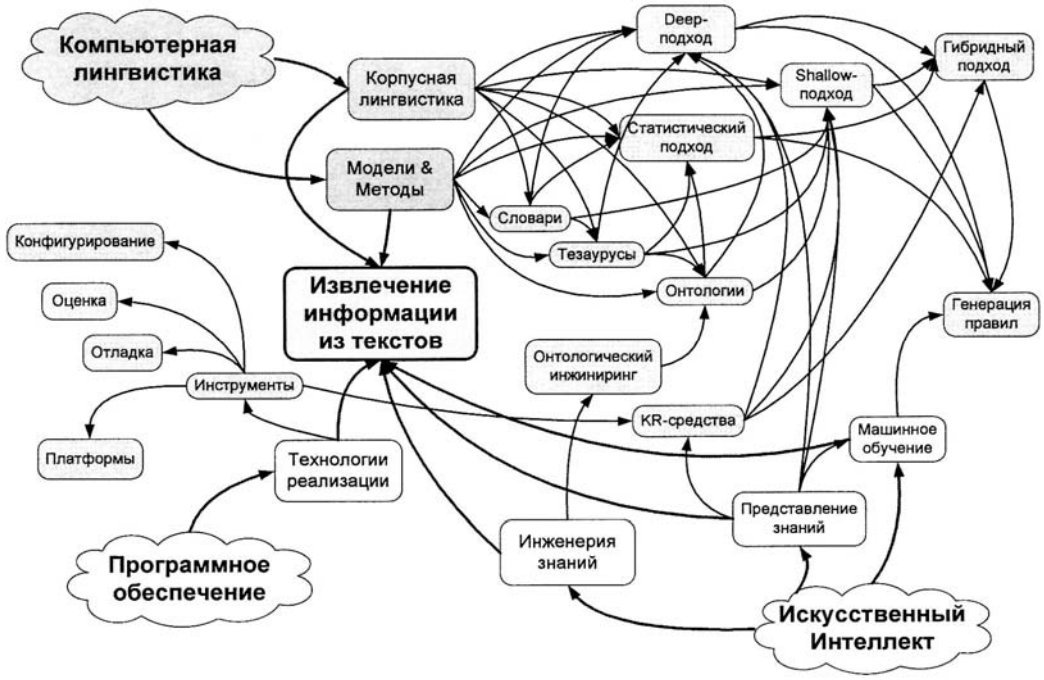

Рис. 1. Структура исследований и разработок в области извлечения информации из текстов

риальное распределение коллективов, представленных на этих конференциях (Рис.2).

Понятно, что интегральные данные дают определенное представление о текущей ситуации, но нуждаются в дальнейшей детализации и анализе, который достаточно трудоемок и не может быть проведен без соответствующих средств автоматизации. Учитывая это, автором в рамках работ по тематике извлечения информации из текстов была разработана специализированная IE-система OntosMiner/SGE (Shadow Groups Extraction), ориентированная на обработку научных статей, и с ее помощью были обработаны труды конференций ДИАЛОГ, КИИ и РОМИП [2, 3]. При этом особое внимание было уделено конференциям серии ДИАЛОГ, как основной профильной конференции по компьютерной лингвистике в России.

Для примера, на Рис.3 представлен геоландшафт конференций серии ДИАЛОГ за 2000-2009 г.г., который показывает, что наиболее активны на конференциях этой серии организации из Москвы (32), Санкт-Петербурга (8), Казани (5) и Новосибирска (4).

Всего из корпуса статей конференций ДИАЛОГ, было выделено объектов типа Person - 1678, Paper - 842, Organization – 204 и Location – 151, между которыми установлены семантические отношения BeAuthor, BeCoauthor, ReferenceTo и симметричное ему отношение ReferencedBy. После проведения автоматической идентификации одинаковых объектов в семантическом пространстве осталось объектов типа Person – 919, Paper – 816, Organization –

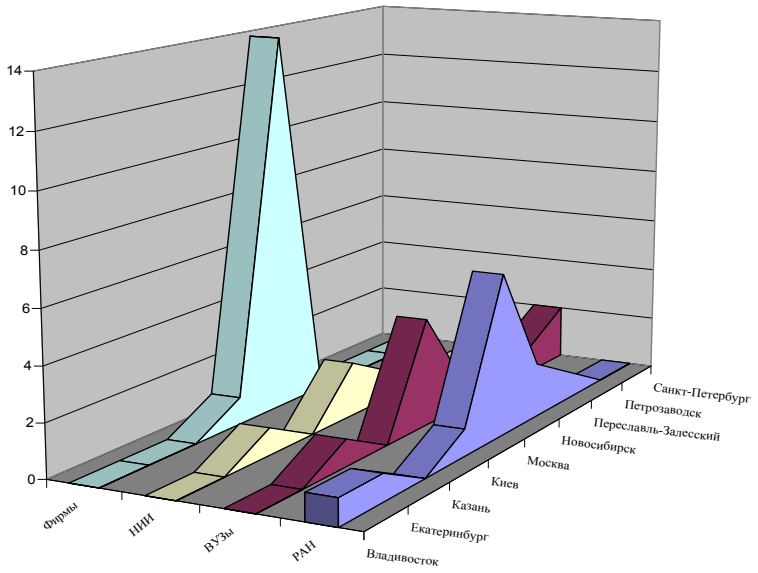

Рис. 2. Организационно-территориальное распределение работ по ЕЯ

105 и Location – 25, а после экспертного их «выравнивания» – объектов типа Person – 854, Paper – 743, Organization – 71 и Location – 13, а также соответствующие семантические отношения между ними (Рис.4).

Какие выводы следуют из анализа диаграммы, представленной на Рис. 4? Во-первых, ее связность при сохранении некоторого числа полностью инкапсулированных авторских коллективов. Во-вторых, появление в диаграмме ярко выраженных «скоплений», что позволяет предположить наличие «скрытых коллективов». Вместе с тем, визуализация всех объектов и связей на одной диаграмме затрудняет ее детальный анализ. Поэтому в этой диаграмме были «скрыты» все объекты типа Paper, Organization

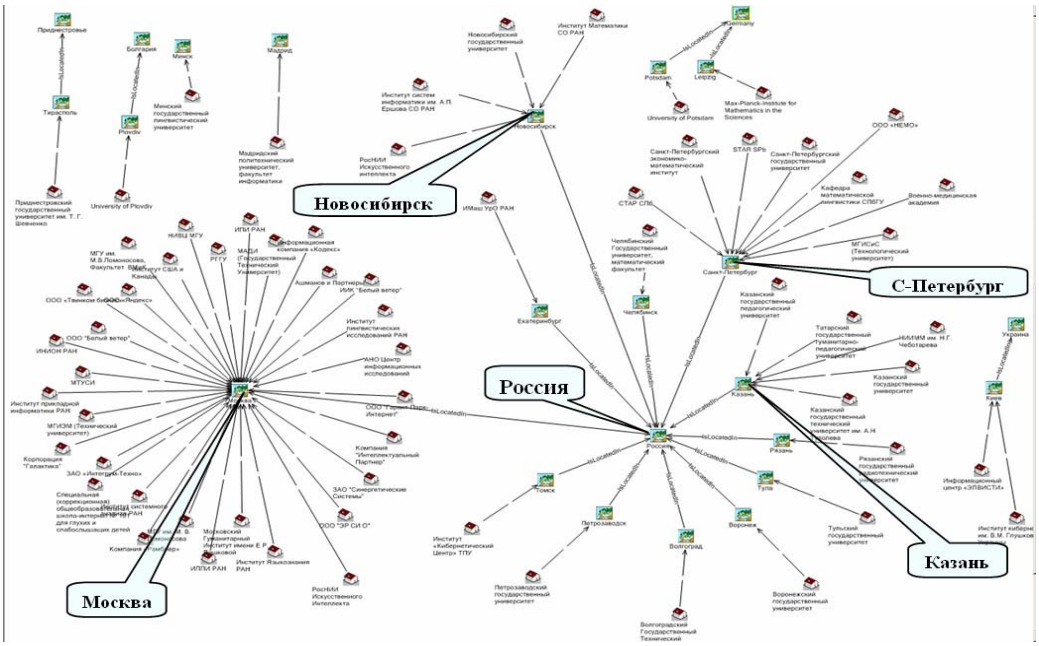

Рис. 3. Геоландшафт конференций серии ДИАЛОГ

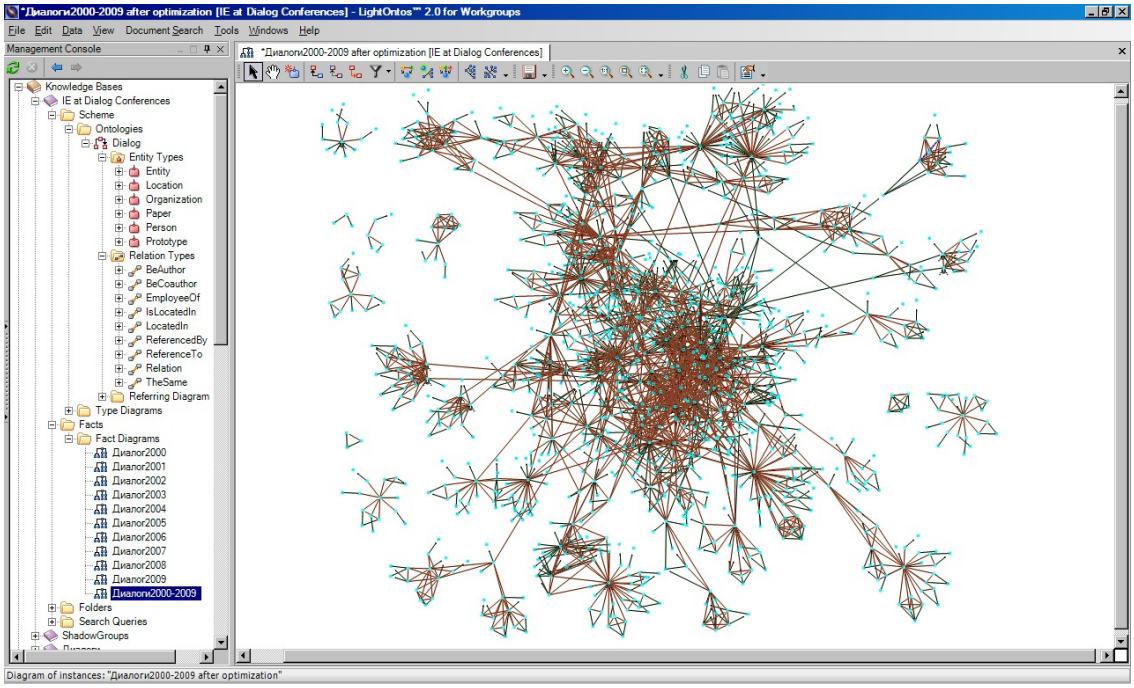

Рис. 4. Общее семантическое пространство конференций серии ДИАЛОГ

и Location, отношения типа ReferencedBy, а также удалены изолированные подграфы. Диаграмма, полученная в результате таких преобразований, представлена на Рис.5 и отражает возможные коллективы и связи между ними. На этой диаграмме выделены наиболее активные кластеры, которые для удобства дальнейшего обсуждения связаны с фамилиями их лидеров.

Как показывает анализ этой диаграммы, все «команды», представленные на конференциях серии ДИАЛОГ, в значительной мере «замкнуты» на себя и при этом подавляющая часть авторов «грешат» автоссылками. Самая многочисленная «команда» представляет несколько тесно связанных между собой организаций и коллективов из Новосибирска (кластер Загорулько [5, 6]), который, при этом, никак не связан с другой новосибирской командой (кластер Загоруйко [7]), которая ссылается только на кластер Кузнецова из ИПИ РАН [8, 9]. Кластер Загорулько ссылается на кластер Ермакова из RCO [10], на кластер Добров-Лукашевич [11], который, в свою очередь, тесно связан общими исследованиями и статьями с кластером Невзоровой [12] (Казань), а взаимными ссылками – с кластером Ермакова [10, 13], где есть ссылка на кластер Антонова [14; 15] и кластер Кузнецова [9], а из кластера Кузнецова - на кластер Ермакова [10]. Из диаграммы на Рис. 5 следует, что наиболее «открыт» миру IE-кластеров нашей страны кластер Большаковой [16, 17], где имеются несколько внешних ссылок [18, 19]. И, наконец, единственный кластер, на который ссылаются три других кластера – это кластер Браславского [20, 21].

Анализ перечисленных выше кластеров и работ российских специалистов, в них представленных, показывает, что в конференциях серии ДИАЛОГ представлены только 3 коллектива (ИПИ РАН, RCO и ВМиК МГУ), основная деятельность которых связана с разработкой и реализацией систем извлечения информации из ЕЯ-текстов, и 2 коллектива (АНО Центр информационных исследований и РосНИИ искусственного интеллекта), которые, скорее, можно отнести к области онтологического инжиниринга и, в частности, использования его результатов для семантического индексирования текстов.

Работы, представленные на конференциях серии РОМИП, находятся вне фокуса данной статьи и практически не вносят ничего нового в геоландшафты российских исследований и разработок в области извлечения информации из текстов. Поэтому читателя, заинтересованного в анализе результатов РОМИП, мы отсылаем к работе [2], где «скрытые» коллективы РОМИП обсуждаются подробнее.

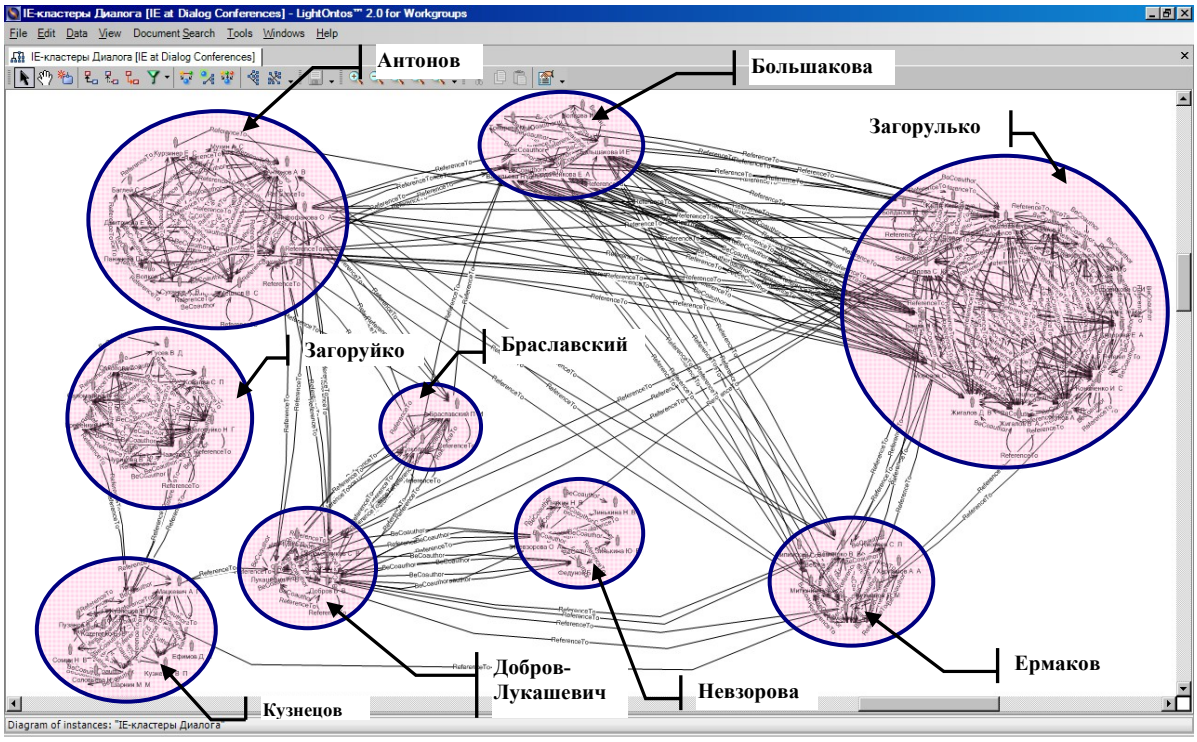

Рис. 5. Диаграмма потенциальных авторитетов и активных IE-кластеров

Что касается конференций серии КИИ, то, в отличие от РОМИП и ДИАЛОГ, работы по ЕЯ составляют лишь 10-20% от общего числа представленных здесь статей, а геоландшафты данной предметной области фиксируют несущественные изменения по сравнению с конференциями серии ДИАЛОГ. Анализ ситуации с исследовательскими коллективами, представленными на конференциях серии КИИ, показывает, что новым (по отношению к конференциям ДИАЛОГ и РОМИП) здесь является кластер авторских коллективов, представляющих проект Ontos [22]. Здесь явных фокусов нет (Рис. 6), но все статьи достаточно сильно связаны ссылками на систему GATE [23], которая использовалась в данном проекте в качестве базовой среды для создания своего инструментария, T.Berners-Lee [24] – «отца» концепции Семантического Веба и двух ведущих специалистов из проекта Ontos – И.В.Ефименко и В.Ф. Хорошевского [25, 26].

Таким образом, по результатам анализа публикаций можно констатировать, что наиболее активными из российских коллективов, работающих в области извлечения информации из текстов, являются «команды» ВМиК МГУ (кластер Большаковой), ЗАО «Авикомп Сервисез» (кластер Ontos), ИПИ РАН (кластер Кузнецова) и RCO (кластер Ермакова). Коллективы АНО ЦИИ (кластер Доброва-Лукашевич) и РосНИИ искусственного интеллекта (кластер Загорулько), как уже отмечалось выше, больше специализируются в области онтологического инжиниринга и использования его результатов для семантического индексирования текстов и потому обсуждаются в отдельной статье.

1.2. Заимка инструментальных средств

Проблема инструментария для создания и развития систем автоматической обработки ЕЯ вообще и систем извлечения информации из ЕЯ-текстов, в частности, всегда была и остается в поле зрения коллективов, претендующих на получение значимых результатов. Вместе с тем, в силу различных причин, так сложилось, что традиция разработки программного инструментария, которая существовала в России в последней четверти XX века [27], была в значительной степени утеряна. И сейчас, особенно в связи с выходом систем обработки ЕЯ на уровень программных продуктов, коллективам, работающим в этой области, приходится восстанавливать эти традиции. Но пока лишь некоторые из команд, создающих прикладные ЕЯ-системы, имеют интересные собственные результаты по инструментальным средствам поддержки разработок, которые и обсуждаются ниже.

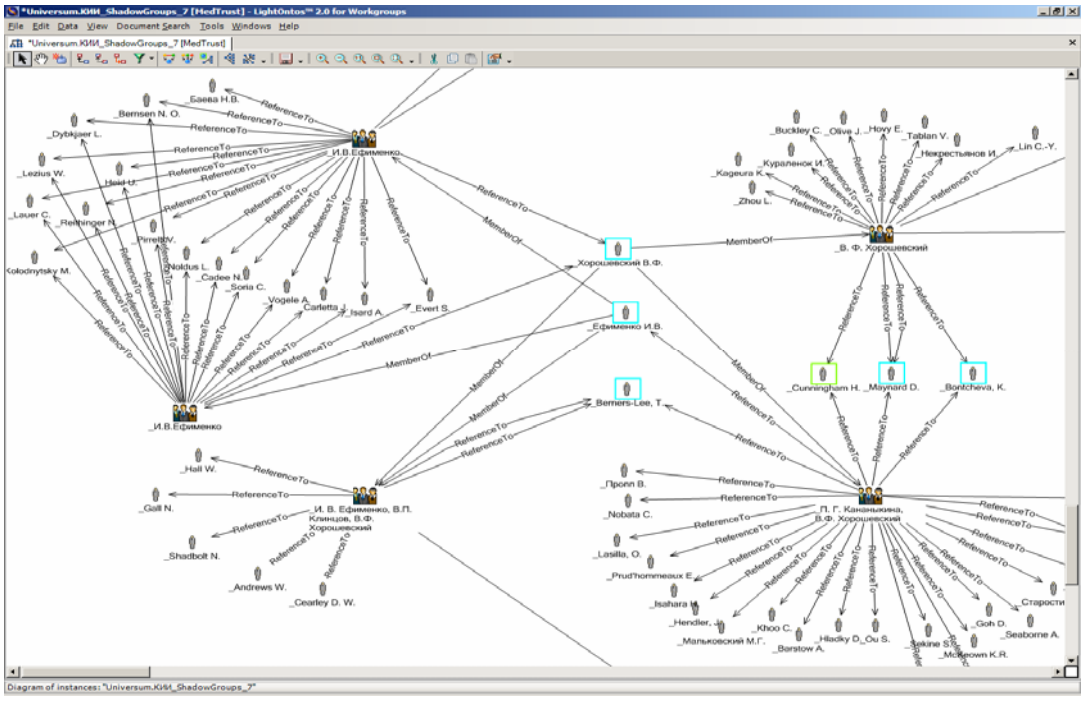

Рис. 6. Кластер авторских коллективов проекта Ontos

1.2.1. Команда ВМиК МГУ (кластер Большаковой)

Как показывает анализ литературы [16,17], группа Е.И. Большаковой занимается исследованиями и разработками в области ЯПЗ, ориентированными на реализацию систем извлечения информации из текстов, уже несколько лет. Результатом этих работ является концепция лексико-синтаксических шаблонов и ЯПЗ SLPL (SyntacticLexico Pattern Language). Как отмечают разработчики, лексико-синтаксические шаблоны – это структурные образцы класса языковых конструкций, отображающие их общие лексические и поверхностно-синтаксические свойства. Формализация таких конструкций в ЯПЗ SLPL в виде шаблонов предполагает определение множества входящих в них лексем и их возможных грамматических форм, а также выявление необходимых синтаксических условий.

Согласно работе [17], лексико-синтаксический шаблон состоит из имени и тела. В общем случае тело шаблона определяет последовательность элементов, из которых должна состоять описываемая языковая конструкция, и задает условия грамматического согласования этих элементов. Так, например, шаблон *AN = A N <A=N>* имеет имя *AN*, тело из элементов *A*, *N* и условие согласования *A=N*. Такой шаблон описывает именную группу из прилагательного *A* и существительного *N,* согласованных по всем их морфологическим параметрам (падеж, число, род).

Основными элементами шаблона являются элемент-строка и элемент-слово. Элемент-строка позволяет описать конкретную словоформу как символьную строку. Элемент-слово соответствует отдельному слову описываемой языковой конструкции, для которого, в общем случае, указываются часть речи, конкретная лексема, определяющая множество всех словоформ слова и значения морфологических параметров слова, сужающие множество допустимых словоформ. Например, элемент-слово *V<пониматься; t=pres, p=3, m=ind>* специфицирует в ЯПЗ SLPL глагол «пониматься» в формах настоящего времени 3 лица изъявительного наклонения. При задании элемента-слова конкретная лексема и значения морфологических параметров могут быть не указаны, что позволяет задать любую словоформу данной лексемы (например, *N<файл>*) или же произвольное слово определенной части речи с нужными грамматическими характеристиками (например, *A<; c=ins, n=sing>* задает любое прилагательное в форме творительного падежа единственного числа).

В общем случае в шаблон могут входить как несколько элементов-слов разных частей речи, так и несколько разных слов одной части речи, а для их различения предлагается использовать числовые индексы. Например, шаблон *NN = N1 N2<; c=gen>* включает два различных существительных *N1* и *N2*, причем второе из них – в родительном падеже.

Условия согласования задаются после описания всех элементов шаблона в виде равенства значений согласуемых морфологических признаков. Так, например, с помощью шаблона *PnV = Pn V <Pn.n=V.n, Pn.g=V.g>* описываются согласованные (в числе и роде) пары из местоимения *Pn* и глагола *V*.

Повторение элементов в шаблоне задается с помощью фигурных скобок, где указываются элементы, которые могут встречаться в тексте несколько раз подряд. Так, например, конструкция *{N<; c=gen>}* задает цепочку из идущих подряд существительных в родительном падеже. Если известны ограничения на количество одинаковых элементов, то их тоже можно указать в шаблоне. Так, запись *{A}<1,3> N* задает последовательность из одного, двух или трех прилагательных и существительного.

Язык LSPL позволяет включать в шаблон опциональные элементы (в квадратных скобках) и альтернативные варианты («|»). К примеру, шаблон *AP = A | Pa* описывает понятие адъектива, т.е. прилагательного *A* или причастия *Pa*.

Лексико-синтаксический шаблон может включать параметры, которые записываются в круглых скобках после всех его элементов и фиксируют те или иные неконкретизированные (т.е. не имеющие значения) морфологические параметры его элементов. Например, параметрами шаблона *AAN = A1 A2 N <A1=A2=N> (N)* являются все морфологические характеристики элемента-слова *N*. В качестве элемента шаблона может быть использован другой, ранее описанный шаблон, что задается именем используемого шаблона и последующими конкретизациями его параметров.

Так, шаблон *TD17 = “далее” “–” NG<; c=nom>* описывает языковые фразы, в которых после слова “далее” через тире идет именная группа *NG* в именительном падеже. В свою

Табл. 2

Шаблон	Пример фразы
TD2 = NG1<; c=ins> V<называться; t=pres, p=3, m=ind> NG2<; c=nom> [PaG] <NG1.n=V.n=NG2.n, PaG=NG2>	*Трансформационным признаком называется приоритетный признак, выделяющий некоторые именные группы в предложении*
TD6 = NG1<; c=acc> ["мы"] "будем" "называть" NG2<; c=ins> <NG1.n=NG2.n>	*Поэтому эту операцию будем называть правилом генерализации примеров*
TD25 = "под" NG1<; c=ins> V<пониматься; t=pres,p=3, m=ind> NG2<;c=nom> <NG1.n=NG2.n>	*…под синтаксемой понимается такое дерево, в корне которого стоит существительное …*
TD18 = NG "(далее"["–"]Ab<;c=nom> ")"	*… все концепты области-источника (далее ОИ), …*
AD1 = NG1<;c=nom> Pa<разработанный; f=short> "в" "целях" NG2<; c=gen> <NG1.n=Pa.n, NG1.g=Pa.g>	*Методика планирования себестоимости услуг разработана в целях обеспечения единства состава и классификации затрат…*

очередь шаблон *NG = {A1} N1 {N<;c=gen>} < A1=N1> (N1)* состоит из существительного *N1* (главного слова), последовательности согласованных с ним прилагательных *{A1}* и цепочки существительных в родительном падеже *{N<;c=gen>}*. Некоторые более сложные примеры шаблонов представлены в Табл.2, заимствованной из работы [17].

В целом ЯПЗ SLPL производит серьезное впечатление и, по мнению его авторов, отвечает интуиции лингвистов, специфицирующих шаблоны различных языковых конструкций. Вместе с тем, в языке не представлены средства спецификации управления выполнением SLPL-программ (порядок применения шаблонов, разрешение многозначностей и т.п.). В опубликованных работах группы нет информации и о том, реализован ли этот ЯПЗ, а если реализован, то каким образом и с какой эффективностью, что позволяет высказать предположение, что SLPL используется в данной исследовательской группе только как общий язык спецификаций.

1.2.2. Команда ЗАО «Авикомп Сервисез» (кластер Ontos)

При разработке технологий обработки неструктурированной информации команда Ontos из ЗАО «Авикомп Сервисез» сориентировалась на хорошо известную в Европе и США платформу GATE [23]. Выбор устойчивой базисной платформы обеспечил коллективу определенные стартовые преимущества, а наличие исходных текстов платформы позволило развернуть собственные работы по ее расширению и модификациям. Все инструментальные разработки команды Ontos концентрируются вокруг создания новых инструментальных компонент самой среды GATE, новых функционалов и технологических компонент, которые интегрированы в эту среду, а также вокруг развития базового языка представления знаний Jape платформы GATE.

К новым инструментальным компонентам среды GATE, разработанным командой Ontos, которые признаны мировым сообществом и вошли в релиз GATE 3.0, относится интерактивный графический отладчик Jape-программ [28, 29]. В рамках создания новых функционалов был разработан спектр компонент, поддерживающих обработку текстов на русском, французском и немецком языках. В частности, реализована русская морфология и система поддержки словарей Dix (Рис. 7), реализованы и интегрированы в GATE специальные Wraper-ы для свободно распространяемых морфологий французского и немецкого языков.

Одновременно велись и ведутся работы по технологическим компонентам инструментария Ontos. В частности, разработан и реализован инструментарий OntoDix, обеспечивающий проектирование предметных и лексических онтологий и отображение их на онтологические словари.

Большой куст инструментальных работ команды Ontos связан с развитием базового языка представления знаний Jape для платформы GATE. Так, уже в 2004 году был разработан новый ЯПЗ Jape+ и компилятор с этого языка [29], которые до последнего времени были базовыми для реализации всех лингвистических процессоров семейства OntosMiner. Язык Jape+ создавался в рамках концепции семейства языков CPSL (Common Pattern Specification Language) [30] и унаследовал наиболее интересные качества таких языков как CPSL, XTDL, Briefs, с одной стороны и простоту базового

ЯПЗ Jape платформы GATE – с другой. При этом ЯПЗ Jape+ существенно мощнее своего «родителя» в части выразительности образцов в левых частях правил, что особенно важно для обработки таких высокофлективных языков, как русский и немецкий, и расширяет набор стратегий управления выводом языка Jape. Так, например, обработку именных групп в соответствии с грамматикой

NP → Adj Noun | Adj NP
Noun → {существительные с морфологическими тэгами}
Adj → {прилагательные с морфологическими тэгами}

можно специфицировать следующим фрагментом Jape + программы:

Как следует из приведенного фрагмента, в левой части правила NounPhrase используются переменные :gend, :nmb и :case для спецификации согласования прилагательных и существительного между собой в роде, числе и падеже в пределах именной группы, а в фазе NounPhraseProc – новый тип стратегии управления выводом brappelt, который является обобщением классических механизмов brill и appelt.

В настоящее время команда Ontos переходит к новому языку представления знаний с рабочим названием Jape-4, который является дальнейшим расширением ЯПЗ Jape+ за счет введения дополнительных средств описания образцов, последовательного отказа от включающего языка Java в правых частях правил

```
Phase: NounPhraseProc
Input: Token Morph
Options: control = brappelt

Rule: NounPhrase
(
  ({ Morph.POS == "A",
     (Morph.GEND != null):gend, (Morph.NMB != null): nmb, (Morph.CASE != null):case }):head
  (
        ({Morph.POS == "A",
            Morph.GEND == :gend, Morph.NMB == :nmb, Morph.CASE == :case }):tail
  )*
  ({Morph.POS == "N", Morph.GEND == :gend, Morph. NMB == :nmb, Morph.CASE == :case }):core
):NounPhrase
-->
```

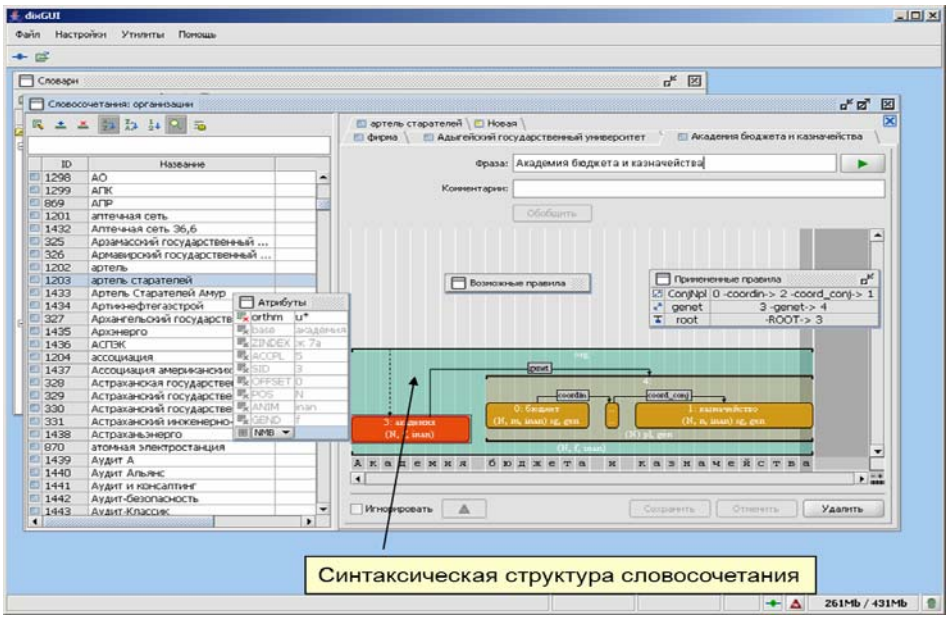

Рис. 7. Экранная форма пополнения терминологического словаря в системе Dix

в пользу операторов высокого уровня для работы с аннотациями, а также за счет использования современных средств повторного использования и комплексирования программных компонент на уровне входного языка. Следует отметить, что в процессе этих работ постепенно происходит отход и от базовой структуры данных платформы GATE – аннотаций в пользу более мощных многомерных структур – тринотаций [31], что вместе с новыми языковыми средствами должно поддержать реализацию новых методов анализа ЕЯ.

Таким образом, можно констатировать, что команда Ontos ведет серьезные собственные разработки в области инструментальной поддержки процессов создания систем обработки ЕЯ, а их результаты использует для проектирования и реализации систем извлечения информации из ЕЯ-текстов.

1.2.3. Команда ИПИ РАН (кластер Кузнецова)

Исследовательская группа И.П. Кузнецова из ИПИ РАН работает в области систем представления знаний, ориентированных на обработку ЕЯ-текстов, уже около 15 лет. Особенность исследований этой группы состоит в использовании логико-аналитического подхода [8], а научно-техническую базу разработок составляют расширенные семантические сети (РСС) и методики представления сложных видов знаний [32], инструментальная среда ДЕКЛ для представления и обработки знаний [9], онтологии в формате РСС [33].

Предметные и лингвистические знания представляются в работах этой группы в нотации семантических сетей, дополненных средствами представления событийных компонент и комплексных связей – РСС. В общем случае РСС состоит из элементарных фрагментов, имеющих произвольное количество аргументных мест и представляющих свойства, отношения, события, действия или множества фрагментов. В простейшем случае фрагменты имеют вид N-местных предикатов (например, *DATA (7,JANUARY,2002)* – фрагмент, представляющий дату) и в этом качестве хорошо коррелируют с фактами языка Пролог [34], хотя автор формализма с этим утверждением не согласен. По утверждению И.П. Кузнецова [8], в общем случае фрагмент – это сложная конструкция, которая выводит формализм за рамки типовых предикатов логики и первого, и второго порядков за счет следующих расширений. Во-первых, в синтаксисе фрагментов используется именование на уровне внутрисистемных кодов – целых чисел с суффиксом плюс, когда вводится новый код, или минус, когда используется уже введенный код. Так, например, в совокупности фрагментов *SUB(MAN,1+) NAME(JOHN,1-)* коды 1+ и 1- представляют одного и того же человека по имени JOHN. Во-вторых, в рамках формализма РСС вводится специальный код фрагмента, соответствующий всей представленной во фрагменте информации. Например, во фрагменте *ОРГ(MORTGAGE,ELECTRONIC,REGISTRATION, SYSTEM,INC./3+)* код 3+ именует (представляет) всю организацию. При этом введенные ранее коды могут стоять на аргументных местах других фрагментов. Так, например, фрагменты

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

представляют, что человек *SHALMAR REESE DANIEL* (ему сопоставлены коды 2+, 2-) передал *(GIVE)* данные организации *ОРГ(MORTGAGE, ELECTRONIC,REGISTRATION,SYSTEM,INC./3+)*, которой сопоставлены коды 3+, 3-. Так коды фрагментов используются для представления комплексной информации и различных видов связей, поскольку РСС ориентированы на отображение возможности интеграции множества связанных объектов в один объект.

Формализм РСС используется в работах группы И.П. Кузнецова на всех уровнях – от морфологии до спецификации смысла текста, полученного в результате его обработки. Так, например, результатом работы блока морфологического анализа является РСС пространственной структуры текста, где представлены слова в нормальной форме с их признаками и указанием их последовательности, а последующая обработка текста сводится к преобразованию РСС на основе заданных правил. При этом собственно морфологический анализ сводится к делению слова на части КОРЕНЬ/ОКОНЧАНИЕ или КОРЕНЬ/СУФФИКС/ОКОНЧАНИЕ, а для выделения окончаний используются, например, следующие фрагменты:

M_OKON_S("IES",3,MANY," "/1+) 1-("Y")
M_OKON_S("S",-1,MANY," ")

Первый фрагмент указывает на необходимость отделения от слова трех последних букв и, если это буквы "IES", выполнения замены "IES" на "Y" (например, слово FLIES после «срабатывания» вышеуказанного фрагмента заменится на FLY). Как утверждают разработчики, аналогичным образом в виде фрагментов могут быть представлены и другие морфологические правила.

Для терминологического анализа в данном формализме используются свои фрагменты вида:

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

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

где *<словоN>* может быть отдельным словом, признаком и даже И-ИЛИ графом. При этом фрагменты типа "ИЛИ" представляется конструкциями вида STR_OR(...), где перечисляются факультативные слова или их признаки, а фрагменты типа "И" – конструкциями вида STR_AND(...), где предполагается обязательность слов с указанными признаками. Для терминов может быть задан как допустимый (слова или их признаки, стоящие слева и справа), так и недопустимый (слова или их признаки, которых не должно быть слева или справа) контекст.

Для представления синонимов и сокращений используются многоместные фрагменты вида *SYNON(<результ.слово>,<исх.слово> ... <исх.слово>)*. Так, например, фрагмент *SYNON(GRAPH, DIAGRAMM)* обеспечивает замену слова *DIAGRAMM* на слово *GRAPH*, а фрагмент *SYNON(CORP., CORPORATION)* фиксирует сокращение *CORP*. от слова *CORPORATION*.

Фрагменты используются и в процессе синтактико-семантического анализа, где по признакам и контексту выделяются информационные или значимые объекты (например, ФИО людей, адреса, организации, номера машин и др.) и для каждого выявленного значимого объекта в документе находится связанная информация (например, для физических лиц это год рождения, пол, адрес и др.). Для выполнения таких преобразований применяются фрагменты, которые называются контекстными правилами и имеют следующий вид:

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

где конструкции вида *<словоN>* могут быть отдельными словами, признаками и И-ИЛИ графами.

Для контекстных правил указывается, с какой позиции начинать их применение, а также допустимый или недопустимый контекст, включая информацию о том, слова с какими признаками не должны стоять в той или иной позиции, что обеспечивает дифференцированное применение правил. При этом контекстные правила применяются в определенной последовательности: вначале выделяются объекты, затем их признаки, словосочетания и, наконец, глагольные формы, а по мере применения таких правил строится РСС как содержательный портрет документа.

Последовательность применения контекстных правил задается с помощью уровней, определяющих порядок применения правил:

LEVEL(LEVEL_1, LEVEL_2, ……..)
LEVEL_1(MORF_ENG) {= Выявление частей речи англ. слов =}
……
LEVEL_2(TTT_1,TTT_2,TTT_3,TTT_4) {= Выделение лиц =}
……

Из доступных спецификаций определения порядка применения правил не вполне ясно, возможны ли вложенные определения порядка, но, как представляется, такие конструкции должны обрабатываться.

Оценивая формализм РСС в целом, можно предположить, что все основные конструкции ЯПЗ ДЕКЛ [9] могут быть реализованы средствами современных версий языка Пролог [35]. Что же касается интуитивной прозрачности рассмотренных конструкций, то, на наш взгляд, она ниже, чем у обсуждавшихся ЯПЗ SLPL и Jape+.

1.2.4. Команда RCO (кластер Ермакова)

Судя по публикациям [36 и др.] и информации, представленной на сайте компании RCO (www.rco.ru), комплексным инструментарием для разработки информационно-поисковых и аналитических систем, требующих лингвистического анализа текстов на русском языке, в данном случае является пакет RCO Fact Extractor SDK.

Ядро пакета представлено библиотекой RCO FX Ru, с помощью которой осуществляется полный синтактико-семантический разбор русских текстов. Библиотека выделяет различные классы сущностей, упомянутых в тексте (персоны, организации, география, предметы, дей-

ствия, атрибуты и др.), и строит сеть отношений, связывающих эти сущности, а также предоставляет всю грамматическую информацию о составляющих текста. Средства библиотеки также обеспечивают семантическую интерпретацию результатов разбора текста – поиск описаний ситуаций, удовлетворяющих заданным семантическим шаблонам.

В состав лингвистического обеспечения пакета, помимо общих словарей и правил русского языка, входят правила выделения специальных объектов (дат, адресов, документов, телефонов, денежных сумм, марок автомобилей и пр.), шаблоны для распознавания различных классов событий и фактов (сделок, экономических показателей, конфликтов, биографических фактов и пр.), тональных характеристик объекта (позитив, негатив и др.), высказываний прямой и косвенной речи.

Язык представления знаний RCO, как следует из работы [36], в целом похож на ЯПЗ Jape платформы GATE [37], но отличается от последнего целым рядом дополнений. Так, например, одно из правил выделения географических названий на ЯПЗ RCO:

```
Rule: GeoNameRule
    (
    ({Token.SemanticType == "Geo:Adj"})
    (({Token.Text =| "регион"}) | ({Token.Text =|
    "район"}))
    ):geo_name
    -->
    :geo_name.Token = { Semantic-
    Type="GeoName:GeoName", Text = geo_name.Text }
```

может быть интерпретировано следующим образом:

левая часть правила задает образец для сравнения атрибутов описаний объектов в распознаваемой цепочке (последовательное упоминание в тексте прилагательного с семантическим типом "Geo:Adj" и любого из заданных слов в любых грамматических формах. При этом предполагается, что семантический тип “географического” прилагательного был присвоен объекту при его выделении на предыдущей фазе или в словарном модуле. В результате будут выделены все объекты типа “Воронежский регион” или “Подольский район”);

правая часть правила указывает, что у выделенного объекта атрибуту Token.Semantic Type будет присвоено значение "GeoName: GeoName", а в атрибут Token.Text будет скопирован текст всей распознанной цепочки.

Как показывает анализ приведенного правила, по сравнению со стандартным Jape, в ЯПЗ RCO используется новый тип присваивания (Text = geo_name.Text) и новый тип сравнения значений (Token.Text =| "регион").

В общем случае, ЯПЗ RCO допускает следующие операции сравнения значения атрибута: «as_is» (==); с регулярным выражением (=~); со строкой с учетом морфологии (=|); со строкой без учета регистра (=^); сравнение на признак «меньше или равно»/«больше или равно» (<= / >=), а также поиск значения атрибута Token.Text в словаре-фильтре с указанным именем (=<).

Управление выводом в ЯПЗ RCO отделено от правил и задается для каждой фазы в секции <rule-phase control=…> файла config.xml. При этом имеется два возможных значения атрибута control: appelt (анализируются все цепочки описаний в графе описаний, полученные в результате выделения объектов на предыдущих фазах) и optimal (анализируется единственная цепочка описаний, соответствующая кратчайшему пути в графе описаний).

Отличаются в ЯПЗ RCO, по сравнению со «стандартным» Jape, и правила именования (binding) конструкций. Так, например, в приведенном ниже правиле выделения спичмейкеров метка :getmorph повторяется в левой части и, кроме того, в правой части используется в дескрипторе спецификации значения, что, на наш взгляд, затрудняет реализацию контроля правильности означивания конструкций и может привести к появлению неоднозначностей:

```
Rule: SpeechMakerRule
( ({Morph.IsAnimate=="Animate"}):getmorph |
({Token.SemanticType=="Name:FullName"}):getmorph
):speach_maker  SPEECH_VERB
-->
:speach_maker.Token={SemanticType="Speech:Maker",
Text=:getmorph.Token.Text},
:speach_maker.Morph = {:getmorph.Morph}
```

Приведенное выше правило позволяет выделить всех «производителей» косвенной речи (спичмейкеров), обозначенных в тексте одушевленными существительными или именами собственными, за которыми следует «глагол говорения», специфицированный макросом SPEECH_VERB.

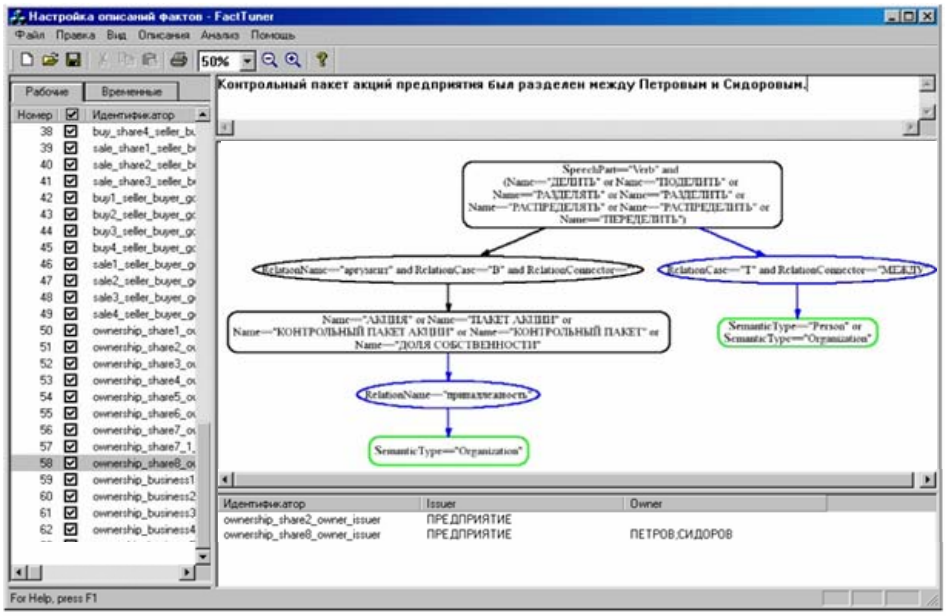

Рис. 8. Интерфейс пользователя инструментальной компоненты RCO Fact Tuner

Имеются и другие отличия ЯПЗ RCO от языка Jape, но в целом он соответствует классу языков семейства CPSL[30].

Из документации к пакету RCO Fact Extractor SDK следует, что команда RCO ориентируется на классические методы пакетной разработки. Вместе с тем, в составе пакета имеется специальная инструментальная компонента RCO Fact Tuner (Рис.8), которая предназначена для создания и настройки семантических шаблонов - описаний фреймов, используемых для выделения в тексте ситуаций и их участников.

Как следует из графического представления семантического шаблона, представленного на Рис.8, в качестве модели здесь используется модификация падежных грамматик Филлмора [38], но из документации неясно, генерируются ли специальные правила выделения из текста ситуаций и их участников по графическому представлению шаблона или формальное представление шаблона интерпретируется специальным «движком».

Помимо рассмотренных инструментальных средств, в команде RCO используются и специальные средства ведения словарей, средства спецификации конфигураций и средства отладки разрабатываемых систем извлечения информации из текстов. Оценивая инструментарий RCO в целом, можно констатировать, что данная команда уделяет этому направлению работ серьезное внимание и даже выводит свой инструментарий на уровень отдельных продуктов компании.

1.3. Поляна извлечения информации из ЕЯ-текстов

В России исследования и разработки в области компьютерной лингвистики имеют давние традиции [39-43], а новый всплеск работ в этой области начался уже более 10 лет тому назад, поскольку в системах семантического аннотирования документов, которые активно востребованы практикой в настоящее время, «узким горлышком» была и остается автоматическая обработка естественного языка.

Учитывая вышесказанное, а также отсутствие общепризнанных моделей и систем автоматической обработки произвольных ЕЯ-текстов, в России, как и во всем мире, активизировалось направление исследований и прикладных разработок, связанных с извлечением информации из текстов на естественных языках. Сейчас в этой области наиболее активными, на наш взгляд, являются коллективы из Москвы [8, 17, 22, 36]. При этом интересно, что часто такие коллективы формируются из специалистов научных институтов РАН и коммерческих организаций, работающих на рынке новых информационных технологий.

Ниже работы этих коллективов в области извлечения информации из ЕЯ-текстов обсуждаются более подробно.

1.3.1. Кластер Большаковой

Как указывалось выше, команда Е.И. Большаковой начала свои работы в области извлечения информации из текстов с разработки своей концепции и поддерживающего ее инструментария. Из практически значимых работ этого коллектива можно отметить подход к автоматическому распознаванию дискурсивной структуры научно-технических текстов и, как следствие, разработку системы извлечения информации о новых результатах, представленных в научных публикациях.

С точки зрения целей настоящего обзора особый интерес в рамках данного подхода представляет исследование структуры научного дискурса, который строится как взаимосвязанная последовательность дискурсивных приемов описания и аргументирования, реализуемых в соответствующих сегментах текста и помечаемых определенными дискурсивными маркерами [44]. Такой подход послужил основой рабочей гипотезы, согласно которой задача автоматического распознавания дискурсивной структуры научного текста может быть решена на основе поверхностного синтаксического анализа текста и лексикона дискурсивных слов. При этом результатом обработки является дискурсивно-композиционная схема текста.

Типичными действиями и операциями научного дискурса здесь являются обоснование вывода, выдвижение гипотезы, введение термина и понятия, приведение фактов и доказательств, подведение итогов и т.п., причем операции эти вводятся автором научного текста и эксплицитно помечаются при помощи разнообразных дискурсивных слов и выражений [45]. При этом маркерами мыслительных операций чаще всего служат ментальные перформативные высказывания (например, особо подчеркнем, далее мы докажем и др.), включающих достаточно широкий круг ментальных перформативных глаголов (выразим, учтем, рассмотрим и т.п.), которые не только помечают, но и квалифицируют соответствующий шаг рассуждения и выстраивают содержание текста в форме рассуждения. Кроме ментальных перформативов к дискурсивным словам относятся «сигналы очередности и логической последовательности» [46], коннекторы и другие выражения, организующие структуру текста-рассуждения, среди которых могут быть метатекстовые операторы (например, подчеркивается, что, по мнению автора и др.). И, наконец, переход от одной мысли к другой в научном тексте осуществляется не только при помощи дискурсивных слов, но с использованием так называемых общенаучных переменных [47] – абстрактных существительных, фиксирующих сам аппарат научно-познавательной деятельности (анализ, гипотеза, проблема, аргумент, следствие, идея, понятие, процедура, модель и др.).

Как представляется, основными научно-техническими результатами исследований команды Е.И. Большаковой в области моделей и методов извлечения информации из текстов являются выделение типовых приемов научного дискурса, соответствующих операциям научного мышления [48]; разработка словаря общенаучной лексики [49, 50]; спецификация процедур дискурсивного анализа, обеспечивающих формирование дискурсивных характеризаций предложений. С практической точки зрения важно, что получаемая в результате обработки текста дискурсивно-композиционная схема может быть использована для получения аннотаций или рефератов. И можно констатировать, что команда Е.И. Большаковой ведет серьезные исследования в области компьютерной лингвистики, которые могут быть использованы в соответствующих системах извлечения информации из научно-технических текстов, но пока законченных реализаций не имеет.

1.3.2. Кластер Ontos

По сути дела, команда Ontos начала формироваться в 2003 г., когда российской IT-компанией ЗАО «Авикомп Сервисез» и швейцарской компанией Ontos был открыт инновационный проект по прикладным системам извлечения информации из текстов. К этому моменту уже было ясно, что компьютерная обработка естественного языка (ЕЯ) снова вернулась в фокус интересов не только исследовательских коллективов, но IT-индустрии и, вместе с тем, было ясно и то, что компьютерное понимание ЕЯ является сложнейшей научно-технической проблемой, полномасштабное решение которой в ближайшее 10-15 лет не просматривается. В связи с этим в проекте, который получил название OntosMiner, а в последствии развился в комплексный проект Ontos, была поставлена задача разработки прикладных систем извлечения информации из

текстов с характеристиками качества, удовлетворяющими требованиям практики. При этом подход команды Ontos к обработке ЕЯ-текстов отличается от подходов большинства других исследовательских коллективов и коммерческих компаний, прежде всего, в следующем: в данном случае НЕ решается и даже НЕ ставится задача полного и правильного анализа произвольных ЕЯ-текстов, вместо чего ставится задача не пропустить те конструкции, которые могут быть обработаны правильно, и не обрабатывать то, что пока правильно обработано быть не может.

Учитывая вышесказанное, команда Ontos сфокусировалась на технологии проектирования и реализации семейства лингвистических процессоров OntosMiner, основные положения которой сводятся к следующему:

использование предметных онтологий для управления обработкой ЕЯ-текстов;

поддержка стандартов консорциума W3C на спецификацию онтологий (в первую очередь, OWL), спецификацию результатов обработки (XML, OWL, N3) и международных стандартов уровня TREC на этапах обработки текстов (аннотации вне текста);

ориентация на такие архитектуры проектируемых систем, которые обеспечивают гибкое комплексирование разнородных компонент за счет использования стандартов на обмен информацией между ними;

многоплатформенная реализация повторно используемых компонент, интеграция которых в рамках мощной инструментальной среды обеспечивает достаточно быстрое проектирование и реализацию коммерческих систем извлечения информации из ЕЯ-текстов.

Онтологический подход к проектированию систем семейства OntosMiner стимулировал создание собственного инструментария онтологического инжиниринга, обсуждение которого выходит за рамки настоящей работы. Поэтому здесь лишь отметим, что в рамках проекта Ontos была, в частности, разработана система поддержки онтологических словарей OntoDix, где словарные входы до погружения в БЗ обрабатываются диалоговым модулем синтаксического анализа словосочетаний [51]. А результаты обработки всех словарных статей, после «утверждения» пользователем, компилируются в эффективный автомат, подключаемый в качестве словарного ресурса к соответствующим системам семейства OntosMiner на этапе исполнения.

Общая архитектура семейства лингвистических процессорах OntosMiner показана на Рис.9.

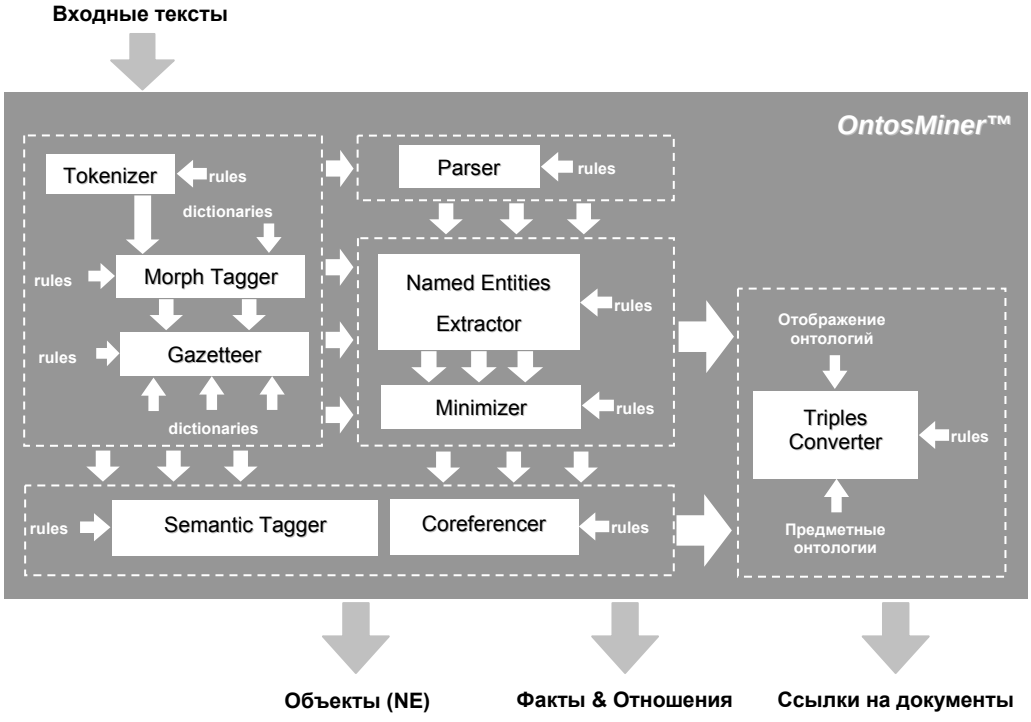

Рис. 9. Общая архитектура систем семейства OntosMiner

В приведенной архитектуре задействованы как классические модули систем типа IE (Tokenizer, MorphTagger, Gazetteer, NE-extractor), так и такие «нетипичные» и/или новые для этого класса систем модули, как Parser, Minimizer, Coreferencer, Semantic Tagger и Triples Converter.

Parser играет в структуре систем семейства OntosMiner роль модуля, обеспечивающего выполнение частичного синтаксического анализа в объеме, продиктованном требованиями практики. Основной задачей модуля Minimizer является разрешение неоднозначностей на выходе модуля NE-extractor. Для этого используется достаточно сложная система весов и правил, которые их учитывают. Как известно [52], автоматический анализ кореферентных и анафорических связей является одной из самых сложных проблем современной прикладной лингвистики, которая до настоящего времени не имеет полного решения. Поэтому в системах семейства OntosMiner модуль Coreferencer решает несколько практически важных, но частных задач. Это обработка наиболее частотных случаев местоименной анафоры и простейших эллипсисов, а также установление кореференций между именованными сущностями с использованием лингвистических правил и статистических методов.

Принципиально новым для IE-систем модулем в системах семейства OntosMiner является модуль Semantic Tagger. На уровне этого модуля из обрабатываемого текста под управлением предметной онтологии с помощью системы семантико-синтаксических правил извлекаются отношения между именованными сущностями и атрибуты этих отношений. Новым для IE-систем модулем является и модуль Triples Converter, поскольку в классических системах его простейшую функцию (вывод результатов обработки) берет на себя инструментальная среда. В данном случае этого недостаточно, и в системах семейства OntosMiner этот модуль не только конвертирует результаты обработки текста во внешнее представление в соответствии со стандартами W3C, но и осуществляет отображение их на внешние предметные онтологии, что позволяет использовать единое внутреннее представление для разных языков (русский, английский, немецкий и др.).

К моменту подготовки настоящей статьи в рамках семейства OntosMiner разработаны несколько процессоров, ориентированных на обработку разных по жанрам текстов на русском, английском и немецком языках. Обобщенные их характеристики представлены в Табл. 3.

В процессе развития проекта Ontos в рамках нескольких заказных и инвестиционных НИОКР были, в частности, разработаны IE-системы для таких предметных областей, как «Политика & Бизнес» (русская и английская версии), «Медицина лекарственных препаратов» (русская версия), «Money Launder» и «HomeLand Security» (английская и немецкая версии), «Нанотех» (русская и английская версии), «Скрытые коллективы» (кросс-языковая версия) и др.

Табл. 3

Модули \ Языки			Русский	Английский	Немецкий
Morph Taggers	обработчики		собственный	собственный	Wrapper для Morphy
	словари		Зализняк-2003	Müller	Morphy
	к-во входов		> 110000	~ 56000	~ 350000
Morph Gazetteers	к-во словарей		85	78	71
	объем		> 1.0M	> 640K	> 1.0M
Parsers	типы конструкций		AdjNP, GenNP, VP	AdjNP, GenNP, VP	VP
	фазы		~ 30	~ 15	~ 35
	правила		~ 100	~ 130	~ 200
Minimizer	правила		~ 150	~ 100	~ 50
Coreferencer	правила		~ 50	~ 30	
Triple Converter	объем			~ 70K	
Content Extraction	NE Extractors	типы NE	~ 21	~ 15	~ 11
		фазы	~ 90	~ 40	~ 60
		правила	> 500	> 200	> 200
	Semantic Taggers	типы отношений	~ 20	~ 10	~ 10
		фазы	~ 60	~ 100	~ 100
		правила	> 300	> 400	> 200

Качество всех процессоров семейства OntosMiner оценивалось на коллекциях документов объемом от сотен до тысяч единиц (для которых был вручную сформирован Gold Standard) в соответствии с методикой TIPSTER [53]. При этом, в зависимости от целей разработки конкретного процессора, предметной области и сложности выделяемых объектов и семантически значимых отношений между ними, точность (P), полнота (R) и F-мера составили: по объектам – P = 0.73-0.95; R = 0.74-0.93; F-measure = 0.75-0.94 и по отношениям – P = 0.67-0.85; R = 0.71-0.83; F-measure = 0.69-0.84, соответственно.

Таким образом, за 7 лет своего существования проект Ontos прошел все стадии – от исследовательских прототипов до коммерчески значимых систем, а сформированная в его рамках команда имеет серьезное реноме в нашей стране и за рубежом.

1.3.3. Кластер Кузнецова

Команда ИПИ РАН под руководством И.П. Кузнецова является одним из типичных для России примеров интеграции исследовательской группы академического института и коммерческой компании, в данном случае ЗАО «Синергетические Системы». По данным сайта компании (www.synsys.ru), эта российская инновационная компания, работающая в области информационных технологий и искусственного интеллекта с 2006 года, среди клиентов которой органы государственного и муниципального управления РФ, корпорации, банки и финансовые группы, крупные библиотеки. Как отмечается в работе [54], совместные работы ИПИ РАН и ЗАО «Синергетические Системы» концентрируются в следующих областях: анализ документов о терроризме на русском и английском языках, обработка сводок криминальных происшествий, анализ обвинительных заключений и справок по уголовным делам, обработка правительственных сообщений, представленных в электронных русскоязычных и англоязычных СМИ, обработка автобиографий и резюме на русском и английском языках.

В этом контексте наибольший интерес, на наш взгляд, представляет лингвистический процессор Semantix [55], предназначенный для автоматической обработки потоков текстов на естественном языке и выделения из них интересующих пользователя объектов и их связей, а также фактов участия объектов в тех или иных действиях или событиях, которые сами рассматриваются как комплексные объекты с их свойствами и связями. В результате обработки на основе каждого документа строится семантическая сеть специального вида (РСС), уже обсуждавшаяся выше. В качестве внешнего представления РСС используется XML, что облегчает последующий автоматический анализ результатов и генерацию досье, обзоров, отчетов и других аналитических документов.

Основные компоненты процессора Semantix: блоки лексического, морфологического и синтактико-семантического анализа; экспертные системы; обратный лингвистический процессор; база лингвистических и экспертных знаний.

Блок лексического и морфологического анализа обеспечивает стандартные для этого класса преобразования и использует специальный набор тематических словарей (словарь стран, регионов России, имен, видов оружия и др.).

Блок синтактико-семантического анализа служит для преобразования пространственной структуры текста в его семантическую структуру и управляется лингвистическими знаниями, которые определяют процесс анализа текста и включают в себя специального вида контекстные правила, обеспечивающие извлечение объектов и связей между ними. По утверждению авторов [56], помимо стандартных задач (извлечение из потока ЕЯ-документов информационных объектов – лиц, организаций, действий, их места и времени и др.; выявления связей между объектами; анализа глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в соответствующих действиях) этот блок выполняет анализ причинно-следственных и временных связей между действиями и событиями.

Специфической компонентой процессора Semantix являются экспертные системы (ЭС), которые на основе результирующей РСС формируют новые знания в виде дополнительных фрагментов РСС. Так, например, при обработке текстов резюме ЭС выявляется область деятельности человека по его автобиографии (в соответствии с заданным классификатором) и оценивается опыт его работы. При анализе криминальных действий ЭС осуществляется отнесение криминального происшествия к определенному типу, выявление характера

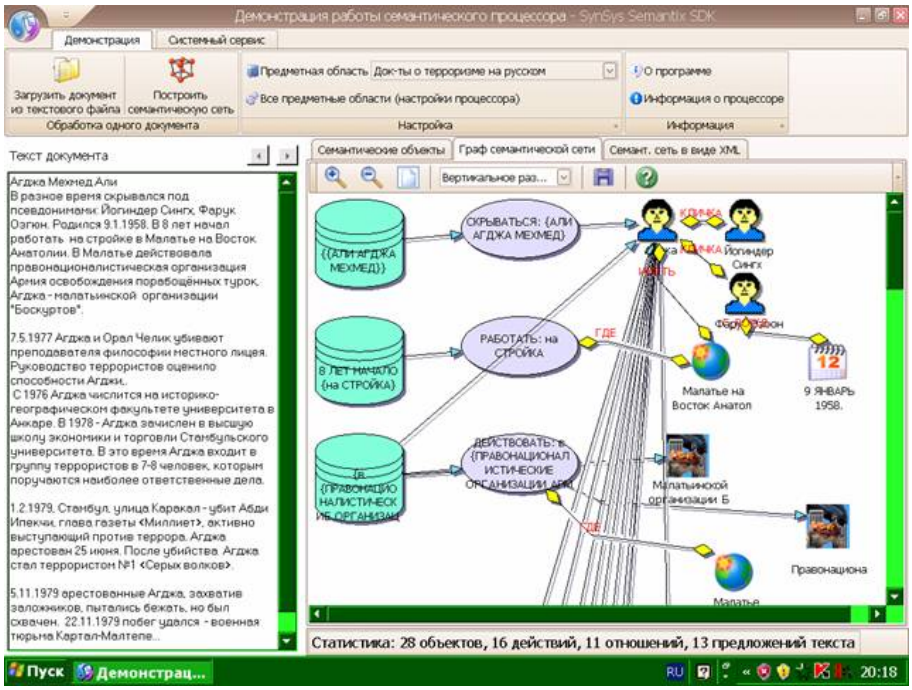

Рис. 10. Графическое представление содержательного портрета документа

преступления, способа его совершения, орудий и т.д. (в соответствии с классификаторами криминальной полиции).

Обратный лингвистический процессор преобразует содержательный портрет документа из РСС в XML и, в случае необходимости, обратное преобразование из XML в РСС.

База лингвистических и экспертных знаний содержит правила анализа текста и экспертных решений во внутреннем представлении, которые определяют работу лингвистического процессора. Semantix имеет несколько таких баз, которые активизируются в зависимости от предметной области и задач пользователя.

Конечным результатом работы процессора Semantix, как указывалось выше, является РСС – содержательный портрет документа. Пример содержательного портрета представлен на Рис.10.

Всего с помощью лингвистического процессора Semantix, как утверждают его разработчики [57], выделяется более 40 типов объектов и связей между ними. При выделении объектов учитываются возможные варианты называния объекта в тексте, а типовые объекты (ФИО, даты, адреса, должности и др.) приводятся к стандартному виду. Кроме того, лингвистическим процессором Semantix выполняется идентификация объектов с учетом их кратких наименований (например, отдельных фамилий или имен с ФИО), анафорических ссылок (указательных и личных местоимений, например, "Этот человек", "Он ..."), определений. При выделении связей между объектами с помощью лингвистического процессора Semantix производится не только глубинный анализ глагольных форм, но и направленный поиск связанных объектов, т.е. восстановление связей, неявно присутствующих в тексте. Для этого, например, организуются специальные процессы, обеспечивающие выделение связи лица с его местом проживания или местом работы, принадлежащим ему автотранспортом и т.д.

К сожалению, разработчики процессора Semantix не приводят данных о характеристиках качества его работы (точность, полнота, F-мера) на достаточно больших и разнородных коллекциях документов, что не позволяет объективно оценить полученные ими результаты, а широта заявленного функционального охвата сложнейших лингвистических феноменов вызывает определенные сомнения в глубине их теоретической проработки и реализации.

1.3.4. Кластер Ермакова

Команда разработчиков RCO имеет репутацию коллектива, давно сосредоточенного на исследованиях и разработках в области обработки текстов на естественном языке. При этом

решения компании ориентированы на работу с большими полнотекстовыми БД, хранилищами данных и системами уровня BI.

Как показывает анализ публикаций команды RCO [13 и др.], основной функционал лингвистических процессоров в данном случае сосредоточен в библиотеке модулей анализа русских текстов RCO Fact Extractor и графического приложения RCO Fact Tuner, уже обсуждавшихся выше. Приложения, построенные на базе библиотеки, как правило, предназначены для аналитической обработки текстов на русском языке и выявления различных фактов, связанных с заданными объектами – персонами и организациями. При этом возможно решение задач из области компьютерной разведки, например, задач автоматического подбора материалов к досье на целевой объект и/или задач мониторинга определенных сторон его активности, освещаемых в СМИ. При этом технологии RCO, как утверждается в публикациях коллектива, позволяют не только найти фрагменты текста, в которых говорилось, например, о поездках персоны, ее встречах, заключении договоров, сделках купли-продажи, но и точно определить все места поездок, визави и контрагентов, наименование товаров и др.

Одним из интересных в контексте настоящей работы результатов команды RCO является разработка системы автоматической оценки высказываний об автомобилях с точки зрения их характеристик и потребительских свойств [58].

Для решения поставленной задачи на базе анализа языкового материала сообщества AUTO_RU "Живого журнала" при помощи средств RCO была сформирована соответствующая предметная онтология, на что, по данным работы [58], было затрачено около 5 чел/дней работы эксперта-автомобилиста (формирование словаря терминов, их классификация по разделам онтологии) и около 50 чел/дней работы лингвиста (отбор и систематизация типовых языковых выражений, создание соответствующих им семантических шаблонов, составление соответствующих словарей оценочных слов, общая настройка и тестирование созданного лингвистического обеспечения). В итоге, из 500000 сообщений “Живого Журнала” было извлечено более 5000 оценок автомобилей, их узлов и характеристик, из которых около 20% (795 «хороших» и 328 «плохих») было привязано к маркам автомобилей. Остальные оценки узлов и характеристик авторам не удалось привязать к конкретным маркам автомобилей в силу различных лингвистических феноменов (например, предложений с синтаксически невыраженным референтом).

Разработанная онтология включает более 700 различных наименований марок автомобилей и фирм-производителей с учетом вариантов написания и известных «народных» названий и более 1200 терминов, структурированных в 24 группы (например, наименования таких узлов автомобиля, как двигатель, коробка передач, ходовая часть и т.п.). Выделено 71 наименование свойств, которые, в свою очередь, были классифицированы на 8 оцениваемых групп (ходовые качества, комфорт, безопасность, надежность и др.), 882 наименования оценок характеристик узлов и свойств, включающих прилагательные, существительные, глаголы и наречия (например, крутой, поломка, глючить, отстойно), в том числе ненормативную лексику, 37 эмоциональных характеристик (например, любить, жалоба, плеваться и т.п.).

Кроме того, в процессе разработки системы было сформировано около 100 семантических шаблонов, описывающих возможные синтаксические связи в предложении между 24 группами терминов из онтологии. Для примера, ниже приведен один из шаблонов, предназначенный для семантической интерпретации фраз, построенных по схеме типа «Размер багажника на Outlander XL вызывает восторг» или «Вид салона Некси приводит в бешенство», заимствованный из работы [58]:

```
if (Name == “ВЫЗЫВАТЬ” or  “ВОЗБУЖДАТЬ” or “ПРИВОДИТЬ” or “ПРИВЕСТИ”){
  if (RelationCase == “И”){
    if (SemanticType == “Artifact:CarUnit” or “Car:Characteristic”){
      if (RelationName == “принадлежность” or (RelationConnector == “В” or “НА” or “У”)){
        if (SemanticType == “Artifact:CarUnit”){
          if (RelationName == “принадлежность” or (RelationConnector == “В” or “НА” or “У”)){
            if (SemanticType == “Artifact:Car” or “Artifact:CarNoName”){
```

```
                действия правой части
              }
            }
          }
        }
      }
    }else{
      if (RelationCase == “” and (RelationConnector == “” or “В”)){
        if (SpeechPart == “Noun”){
          действия правой части
        }
      }
    }
  }
```

Шаблон фиксирует лексико-грамматические ограничения на искомую конфигурацию связей между словами в тексте, которые определяются синтаксическим анализатором. В частности, в шаблоне указываются ограничения на конкретные слова (Name=" ВЫЗЫВАТЬ"), части речи (SpeechPart="Noun") или семантические разряды слов (SemanticType="Artifact:CarUnit"), а также ограничения на синтактико-семантические связи между словами (RelationName == "принадлежность"), предлог (RelationConnector == "НА"), семантический падеж (RelationCase=="И"). Окончательно такой шаблон параметризуется множеством конкретных слов из онтологии. Так, названиями эмоций параметризуется элемент шаблона с ограничениями SpeechPart="Noun", а названиями узлов автомобиля и их характеристик – элементы шаблона с ограничениями SemanticType="Artifact:CarUnit" и SemanticType="Artifact:CarCharacteristic".

В результате экспериментов, как отмечается в работе [58], достигнутая точность извлечения составила 84%, а полнота – около 20%. Анализ ошибок, проведенный авторами, показал, что точность и полнота могут быть повышены за счет дальнейшей доработки онтологии незначительно. В частности, принципиальную проблему дальнейшего повышения точности и полноты представляет интерпретация фраз, содержащих отрицание, выраженное сложным образом (например, логически: «Я нигде, никогда не писал, что Мерседес, БМВ и Ягуар имеют высокую надежность»; «А вообще чтоб в такую погоду на 60° Лексус занесло - это бред!» и т.п.; метафорически: «За 7 лет от надежности Сивика остались одни воспоминания») и с помощью других сложных лингвистических конструкций, что полностью совпадает с собственным опытом автора по автоматическому определению тональности текстов.

Несмотря на относительно невысокие показатели точности и, особенно, полноты вывод автора работы [58] о том, что в проведенном эксперименте удалось показать практическую возможность извлечения из социальных сетей Интернета полезных знаний представляется обоснованным с учетом большого количества грамматических и орфографических ошибок, особого стиля этого "жанра" текстов, большого количества сленговой лексики, а окончательно ожидаемая автором точность в районе 90% вселяет надежду на возможность извлечения знаний из Интернет-«помоек» с приемлемым качеством, поскольку недостаточная полнота (20%) здесь компенсируется избытком информации.

Оценивая потенциал команды RCO в целом, можно констатировать следующее: разработки коллектива оставляют серьезное впечатление, концепция разработки лингвистических процессоров очень близка к Shallow Approach [59], а в качестве языка представления знаний используется собственная разработка ЯПЗ, похожего на язык Jape среды GATE, стандарты W3C поддерживаются. Таким образом, команда RCO является одной из самых продвинутых российских команд, работающих в области извлечения информации из ЕЯ-текстов на русском языке.

1.4. Ярмарка продуктов и сервисов

В предыдущих разделах обсуждался, так сказать, базис извлечения информации из ЕЯ-текстов, представленный в работах российских коллективов и специалистов. В настоящем разделе рассматриваются продукты и сервисы, которые формируют надстройку над этим базисом, которая, собственно, и важна для

пользователей. И если, как отмечалось выше, получить достоверную и сколько-нибудь полную информацию по базису было достаточно трудно, то для надстройки существует другая проблема – информации о системах, претендующих на интеллектуальную обработку текстов, много, но она, как правило, ориентирована на потенциальных покупателей таких систем и потому часто носит рекламный характер.

С учетом вышесказанного, ниже приведены краткие описания семантических продуктов и сервисов, которые, на наш взгляд, интересны в контексте данной работы и представлены на рынке наиболее известными (активными) российскими производителями. При этом автор надеется, что квалифицированный читатель сможет сам определить степень соответствия этих описаний действительности и потому в этом разделе постарается воздерживаться от их оценки.

1.4.1. Система Медиалогия

По информации с сайта (www.mlg.ru) Медиалогия является компанией-разработчиком первой в России автоматической системы анализа и мониторинга СМИ в режиме реального времени, не имеющей аналогов ни в нашей стране, ни за рубежом. Медиалогия поставляет решения для анализа СМИ крупнейшим российским компаниям и госструктурам.

Система Медиалогия – флагманский и, похоже, единственный из известных продуктов компании имеет две части. Это база данных СМИ и аналитический модуль для обработки и анализа сообщений СМИ. Ежедневно обрабатывается порядка 45000 сообщений из 3500 СМИ. Все сообщения поступают в базу данных в круглосуточном режиме по жесткому регламенту и проходят автоматическую обработку с последующим контролем. Обработанные сообщения хранятся в БД СМИ и доступны для поиска, просмотра и анализа. Обработка сообщений автоматизирована на 85%, что позволяет гарантировать необходимый уровень качества обработки и выдерживать жесткие требования по срокам доступности сообщений в системе с момента публикации или выхода в эфир. По утверждению разработчиков, в каждом сообщении выделяются объекты (физические или юридические лица, бренды и географические понятия), причем, в отличие от традиционного контекстного поиска, система "умеет" точно идентифицировать объект, получая дополнительные данные о нем из контекста, например, может "отличить" однофамильцев или компании с похожими названиями, но из разных отраслей. Точность объектного поиска (по данным внутреннего тестирования) составляет 98%.

Для каждого объекта и сообщения в целом определяется ряд параметров. Для объекта это роль в сообщении (главная/второстепенная), наличие прямой речи, связи, характер упоминания (позитив/негатив). Для сообщения – размер статьи или хронометраж, полоса или время выхода в эфир, наличие иллюстраций, экспрессивность заголовка и др. Все эти параметры (метаданные) хранятся в базе вместе с сообщением и позволяют конкретизировать условия поиска, а также строить в режиме online статистические и аналитические отчеты. Результаты представляются в виде списков сообщений, таблиц, карт, диаграмм и графиков. Все данные, представленные в отчетах, можно проверить, кликнув на соответствующую цифру таблицы или точку графика и перейдя к списку сообщений, которые стоят за конкретным результатом.

Как представляется, наиболее интересным в семантических технологиях компании Медиалогия является Индекс Информационного Благоприятствования (ИИБ)™, предназначенный для анализа СМИ с учетом их влиятельности, яркости и характера упоминания. ИИБ применяется пресс-службами и аналитиками для решения задач оперативного отслеживания негатива и потенциальных информационных рисков, круглосуточного мониторинга позитивных и негативных сообщений в условиях информационных войн и кризисов, анализа освещения информационных поводов с учетом влиятельности СМИ и количества перепечаток, оценки изменений качества информационного поля компании в сравнении с конкурентами за определенный период времени и т.п. По информации с сайта ИИБ рассчитывается автоматически с применением технологий лингвистического анализа по методике, разработанной компанией совместно с учеными-математиками и аналитиками масс-медиа.

Как уже отмечалось выше, автоматический анализ тональностей является одной из самых сложных задач обработки ЕЯ-текстов. Авторы системы утверждают, что им удалось решить эту задачу с приемлемым для практики качест-

Рис. 11. Оценка медиа репутации компании с помощью системы Медиалогия

вом, но, правда, отмечают, что оценки, автоматически проставленные лингвистическим модулем, проверяются редакторами компании и только после этого распространяются на весь шлейф перепечаток сообщений на ту же тему. Тональность в системе Медиалогия определяется наличием в статье экспрессивных выражений по отношению к объекту. Учитывается как тональность автора сообщения, так и комментарии других участников публикации, однако авторская позиция превалирует над остальными. Заметим, что утверждение о том, что такие нюансы, как например, сарказм выделяются автоматически с заявленной точностью и полнотой, автору кажутся, мягко говоря, преувеличением.

Понятно, что если задача автоматического (автоматизированного) определения тональностей решена, можно оценить медиа репутацию компании относительно конкурентного окружения, а также ее изменение с течением времени. На Рис. 11 приведен пример динамики количества сообщений и ИИБ™ компании за некоторый промежуток времени, скопированный с сайта описания продукта Медиалогия, где видно, что всплески интереса СМИ к компании вызваны определенными информационными поводами, положительно и отрицательно влияющими на ее репутацию.

В системе Медиалогия решается и классическая задача извлечения информации из текстов – выделение именованных сущностей типа Люди, Организации и, по-видимому, некоторых других, но ни в публикациях, ни на сайте компании нет информации о том, какие методы для этого используются. С точки зрения настоящей работы интересной является информация о том, что в системе Медиалогия имеется функция выявления связей между объектами (персонами и компаниями). При этом утверждается, что система не только идентифицирует факт наличия связи, но и определяет по контексту ее характер (тип) из следующего набора: «Контакт», «Конфликт», «Отставки и назначения», «Упоминает», «Сделка», «Неоднозначная связь». Правда и в этом случае отсутствует информация о том, какие методы для решения задачи выделения типизированных связей используются и какие характеристики точности и полноты при этом получены.

В целом можно констатировать, что функционалы системы Медиалогия безусловно полезны и впечатляют, система реально используется на практике. Вместе с тем, по косвенным данным, многие функционалы, которые позиционируются на сайте как автоматические, в действительности являются автоматизированными и поддерживаются за счет четкого соблюдения технологии обработки информации людьми-экспертами.

1.4.2. Система «Галактика ZOOM»

Система «Галактика ZOOM» – один из продуктов корпорации «Галактика». Хотя на сайте самой корпорации (http://www.galaktika.ru) явного указания на это нет, такая информация имеется на сайте (http://www.galaktika-zoom.ru).

Там же декларируется, что «Галактика ZOOM» – инструмент для создания хранилища текстовой информации, который обладает уникальными возможностями для проведения эффективного поиска и аналитических исследований и, по мнению независимых экспертов (российских и западных), не имеет аналогов по функциональным возможностям и потенциалу.

Преимущества системы «Галактика ZOOM», по мнению ее авторов, состоят в том, что она обеспечивает уникальные функции формирования «информационного портрета» объектов, проведение сравнительного, динамического и объектного анализа; быструю обработку информации объемом до десятка терабайт; использование страничных механизмов организации данных; масштабируемость, а также имеет понятный графический интерфейс и общепринятый язык запросов. Для руководителя система «Галактика ZOOM» позволяет в режиме реального времени увидеть ситуацию во всех ее проявлениях; для аналитика – оперативно составлять отчеты, описывающие ситуацию и тенденции ее развития в политической, экономической, социальной и других сферах; для маркетолога – исследовать рынок, выявлять источники потенциальных возможностей и контролировать деятельность компаний-конкурентов. Специалистам по PR система «Галактика ZOOM» помогает проводить мониторинг СМИ, отслеживать PR-акции конкурентов, формировать дайджесты, а сотрудникам службы безопасности – выявлять источники угроз, как со стороны внешнего окружения, так и внутри компании и формировать досье на физических и юридических лиц.

Таким образом, если следовать логике рекламного проспекта по системе «Галактика ZOOM», она обеспечивает все потребности пользователей данного сегмента рынка и, по существу, является прямым конкурентом системы Медиалогия.

1.4.3. Продукты компании «ЭР СИ О»

Как известно, ООО «ЭР СИ О» (www.rco.ru), торговая марка RCO, выделилась в самостоятельную компанию в 2006 году. До этого команда RCO работала в компании «Гарант-Парк-Интернет». За последние годы компания участвовала в выполнении нескольких крупных проектов по заказу государственных и коммерческих организаций.

На рынке продуктов, связанных с обработкой естественного языка и семантическими технологиями, RCO занимает устойчивое положение в течение последних 3-5 лет. В настоящее время в ее портфеле имеются продукты, краткое описание которых представлено ниже.

RCO for Oracle, по утверждению компании, единственный на рынке продукт, позволяющий значительно расширить возможности OracleText при работе с базами данных, содержащими документы на русском языке.

RCO for BackOffice позволяет расширить возможности Microsoft BackOffice (MS SharePoint Portal, MS Indexing Service, MS Exchange Server и MS SQL Server) при работе с документами на русском языке, обеспечивая поиск с учетом всех грамматических форм слов на основе морфологического анализа.

Кроме того, компания позиционируется в продуктах RCO Morphology SDK и RCO RCO Morphology Professional SDK, а также в продуктах класса «Инструментарий аналитика» (RCO Fact Extractor и RCO KAOT – наиболее интересных продуктах с точки зрения настоящей работы.

RCO Fact Extractor – персональное приложение для Windows, которое обсуждалось выше.

RCO KAOT – информационно-аналитическая система для работы в локальной сети на базе MS Windows и MS Internet Information Server. Продукт представляет базовое решение для организации автоматизированного рабочего места аналитика и может быть использован для решения широкого класса задач: от контекстного поиска документов с учетом всех словоформ, синонимов и опечаток до поддержки принятия решений на основе анализа информационных массивов с применением методов искусственного интеллекта.

Оценивая потенциал ООО «ЭР СИ О» в целом, можно констатировать, что в компании имеется достаточно мощная среда разработки лингвистических процессоров, а все продукты компании разбиты на 3 подкласса: решения по русификации систем известных производителей (Oracle, MicroSoft), решения по обработке естественного языка и решения по аналитике на базе использования лингвистического процессора собственной разработки.

1.4.4. Интеллектуальная поисковая система Exactus

Интеллектуальная поисковая система (ИИПС) Exactus (www.exactus.ru), вообще говоря, находится вне основной линии настоящего обзора, а включена в него потому, что это, насколько известно автору, одна из немногих ИПС, где явно декларируется активное использование лингвистических технологий для анализа запросов и индексирования документов [60, 61].

По своей функциональной структуре ИПС Exactus относится к классу метапоисковых систем и вместо создания собственного индекса использует результаты, полученные по запросу от других поисковых машин – Google, Yandex, Rambler, Ask, MSN и Yahoo, которые обрабатываются (как и запрос) лингвистическим процессором Exactus. Таким образом, алгоритм поиска Exactus объединяет статистическую и лингвистическую составляющие. Из статистических характеристик текста Exactus учитывает TF*IDF веса термов и значимость фрагментов текстов (на основе HTML-разметки документов). Лингвистическая составляющая – значения синтаксем (минимальных семантико-синтаксических единиц текста) и их сочетаемость в конкретном предложении [62], что позволяет отбирать только те тексты, в которых семантическое значение синтаксемы совпадает с ее семантическим значением в запросе. Так, например, при обработке запроса «К чему приводит инфляция?» и двух ответов, полученных от других поисковых машин («Инфляция приводит к снижению темпов экономического роста» и «Строительство непроизводственных мегаобъектов приводит к росту инфляции»), для системы Exactus более релевантным является первый документ, так как во втором документе «инфляция» находится в другом семантическом значении.

Методы семантического поиска, реализованные в системе Exactus, основаны на теории коммуникативной грамматики русского языка [62] и теории экстралингвистических семантических отношений [63]. Спецификой системы Exactus является возможность выбора различных стратегий (профилей) поиска, задаваемых пользователем. В системе поддерживаются 4 профиля: «ситуация», «факт», «объект» и «автомат». Профиль «ситуация» описывается объектами предметной области (участниками ситуации) и отношениями между ними. При этом интерес представляют типы отношений между участниками, а не сами участники. Для профиля «факт» (ситуации с фиксированными участниками) важны и участники, и отношения между ними, а для профиля «объект» (участники ситуации) отношения между участниками менее важны, чем характеристики участников. В случае выбора профиля «автомат» система пытается определить, что имел в виду пользователь, и подобрать для поиска соответствующий тип профиля автоматически.

1.4.5. Сервисы компании “Яндекс”

Как известно, Яндекс – крупнейший российский портал, предлагающий пользователям ключевые Интернет-сервисы. Компания активно развивается и в направлениях, связанных с обработкой естественного языка и внедрением в свои сервисы семантических технологий. С точки зрения настоящей работы наиболее интересным является проект ” Яндекс.Новости”, в рамках которого, по утверждению пресс-службы компании, успешно применяется технология извлечения фактов, принадлежащая компании «Яндекс». На ее основе реализованы локальные проекты “Пресс-портреты” в Новостях; “Цитаты” в Новостях и “Новости регионов”.

Исторически первым из Яндекс-сервисов данного типа был сервис «Пресс-портреты», в рамках которого представлены три типа сведений о человеке: свободные определения человека (звания, ученые степени, профессии и др.), послужной список (факты, состоящие из названия организации и должностей, которые человек в ней занимал, занимает или займет в будущем) и цитаты (цитаты человека и цитаты о нем других людей). Каждый факт снабжен текстовой иллюстрацией – фрагментом новостного сообщения, из которого была извлечена информация. Кроме того, для каждого факта можно получить все его упоминания в “Новостях”.

Процесс автоматического составления пресс-портрета по материалам сообщений из ”Новостей” состоит из двух основных этапов: выделение фактов из текста (объекты и отношения между ними) и формирование пресс-портрета (кластеризация фактов, группировка фактов, относящихся к одному человеку).

На первом этапе выделяются типизированные объекты: ФИО, название организации, описание организации, должность, геоимена.

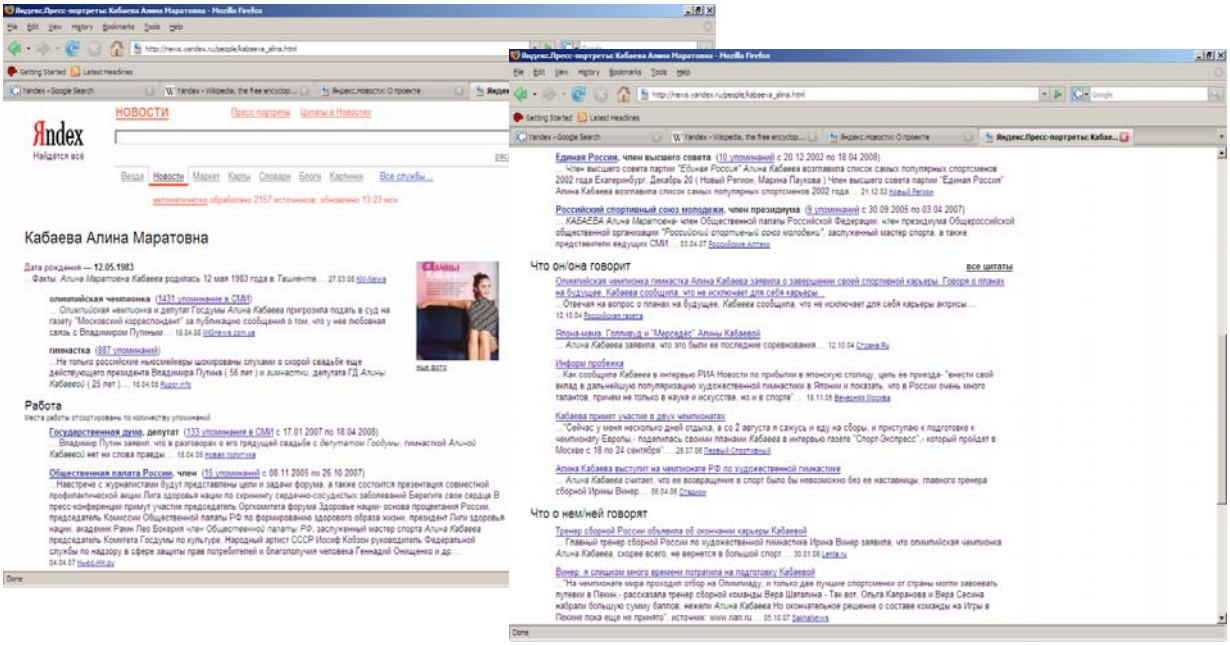

Рис. 12. Экранные формы пресс-портрета Алины Кабаевой

Основной принцип выделения объектов состоит в выделении опорных (ключевых) слов, которые могут быть вершинами обозначающих их синтаксических групп. На следующем шаге распознаются неразрывные цепочки, состоящие из построенных объектов, которые характеризуются порядком их следования, согласованием отдельных элементов или грамматическими характеристиками объектов. Для распознавания цепочек используется набор шаблонных правил, которые позволяют задавать все эти характеристики. Для распознавания отношений в случае неконтактного расположения объектов используется модуль фрагментационного анализа, разбивающий предложение на простые фрагменты и определяющий вершину каждого фрагмента (подлежащее и сказуемое для простых предложений). Кроме того, подключается словарь предикативных вершин (глаголы, причастия, предикативные словосочетания), в котором для каждой вершины описаны все ее актанты (участники ситуации) и способы их выражения в предложении (порядок следования и морфологическая информация). На втором этапе извлеченные факты объединяются в пресс-портрет (Рис. 12).

Сервис Яндекса «Цитаты в Новостях», по сути дела, является побочным эффектом разработки технология извлечения фактов, использованной при построении пресс-портретов, а обобщающим в данном домене является сервис ”Новости”. По утверждению пресс-службы компании, это первая в России служба автоматической обработки и систематизации новостей, которая обновляется в режиме реального времени 7 дней в неделю и 24 часа в сутки и беспристрастно отображает информационную картину дня. Данные для обработки поступают от партнеров службы – ведущих российских и зарубежных СМИ. Каждое поступившее сообщение об актуальном событии сразу же включается в посвященный этому событию сюжет. Сюжеты ранжируются по важности. Отсутствие человеческого вмешательства позволяет объективно формировать сюжеты, помещая рядом сообщения с совершенно разными точками зрения. Таким образом, пользователь получает возможность сравнить, как одно и то же событие отражено в различных СМИ.

Ключевым, с точки зрения технологии, в данном сервисе является понятие сюжета (подборки сообщений СМИ, посвященных одному событию). Формирование сюжетов происходит путем определения текстуальной близости и кластеризации информационных потоков. Ранжирование сюжета зависит, в первую очередь, от его актуальности и количества сообщений в сюжете, а также от интереса пользователей. Отображение сюжета основано на технологии «многодокументной аннотации». Из всех со-

общений сюжета автоматически выделяются наиболее значимые объекты (имена людей, названия организаций, геообъекты, даты и числа), которые, наряду с ключевыми словами сюжета и новостными запросами, определяют выбор текстов для аннотации. Сюжет иллюстрируется фотографиями и картами (если в нем упомянуты какие-то геоимена). От основных действующих лиц сюжета можно сразу перейти на их пресс-портреты, с региона сюжета – на все новости данного региона (интересный «заход» на семантическую навигацию). При выборе сообщений для показа их на первой странице сюжета учитывается положение источника сообщения в общем рейтинге. Рейтинг строится на основе трех показателей: оперативность, эксклюзивность и разнообразность.

Новости на сайте структурируются по темам и географии. При этом выделяются главные новости и новости в блогах. Разработчики утверждают, что имеется возможность увидеть «Тему дня» (совокупность крупных сюжетов, посвященных происходящим параллельно и объединённым единой темой событиям) и «Новость часа», которая определяется путем автоматической оценки всех новостных сюжетов и выяснения, какие из них в настоящее время демонстрируют всплеск активности, и выделения одного – с наиболее сильным и устойчивым ростом.

Оценивая потенциал компании «Яндекс» в сегменте сервисов, при создании которых использовались средства обработки ЕЯ-текстов и семантические технологии, можно отметить, что у компании, как у мощного информационного портала, имеется серьезная аппаратная инфраструктура и постоянно обновляемый и проиндексированный контент, на который опираются все сервисы. Вместе с тем, лингвистические технологии в компании «Яндекс» пока используются в ограниченном объеме, а основная ориентация разработок лежит в сфере статистических методов обработки ЕЯ.

1.4.6. Продукты и сервисы линейки Ontos

Научно-технические результаты команды Ontos уже обсуждались выше, а в данном подразделе кратко рассматриваются некоторые из продуктов линейки Ontos как их «видит» пользователь.

На сайте ЗАО «Авикомп Сервисез» (http://www.avicomp.ru) направление «Семантические технологии» позиционируется как одно из ключевых, а все работы концентрируются на разработке систем семантической обработки ЕЯ-текстов под управлением предметных онтологий, на создании программно-аппаратных платформ для семантизации Интернет-контента и корпоративных хранилищ информации, а также реализации прикладных систем семантизации контента, эффективного поиска и навигации по хранилищам знаний и информационно-аналитических систем нового поколения.

По данным с сайта компании, ЗАО «Авикомп Сервисез» успешно реализует проекты по разработке, внедрению и сопровождению решений на основе семантических технологий в государственных учреждениях и в крупных коммерческих структурах, а также первой из российских IT-компаний открыла для внешнего тестирования через Интернет функционал семантической разметки текстов пользователя и API для разработки внешних приложений.

В результате работы систем извлечения информации из ЕЯ-текстов осуществляется семантизация контента. Так, на Рис.13 для примера приведена страница Google management Team, посвященная Сергею Брину, и экранная форма с результатами ее обработки.

Извлечение информации из текстов с последующим сохранением результатов в семантическом хранилище открывает возможности семантической навигации – перемещения между объектами базы знаний по связям между ними с получением информации о свойствах этих объектов и перемещениям по документах-источникам информации (Рис.14).

Важнейшим отличительным свойством системы семантической навигации является то, что она работает на множестве сайтов близких тематик как одном мега-портале. При этом пользователям семантически проиндексированных ресурсов не надо просматривать множество страниц, полученных поисковыми машинами, чтобы найти всю интересующую их информацию, а достаточно выбрать соответствующий объект на странице любого из проиндексированных сайтов и открыть его карточку, где сразу будет отображена обобщенная информация, собранная со всех страниц обработанных ресурсов.

Типичными примерами семантических сервисов, разработанных командой Ontos, являются семантическое дайжестирование и реферирование. Первый из них позволяет автоматически генерировать отчеты (дайджесты) из фрагментов

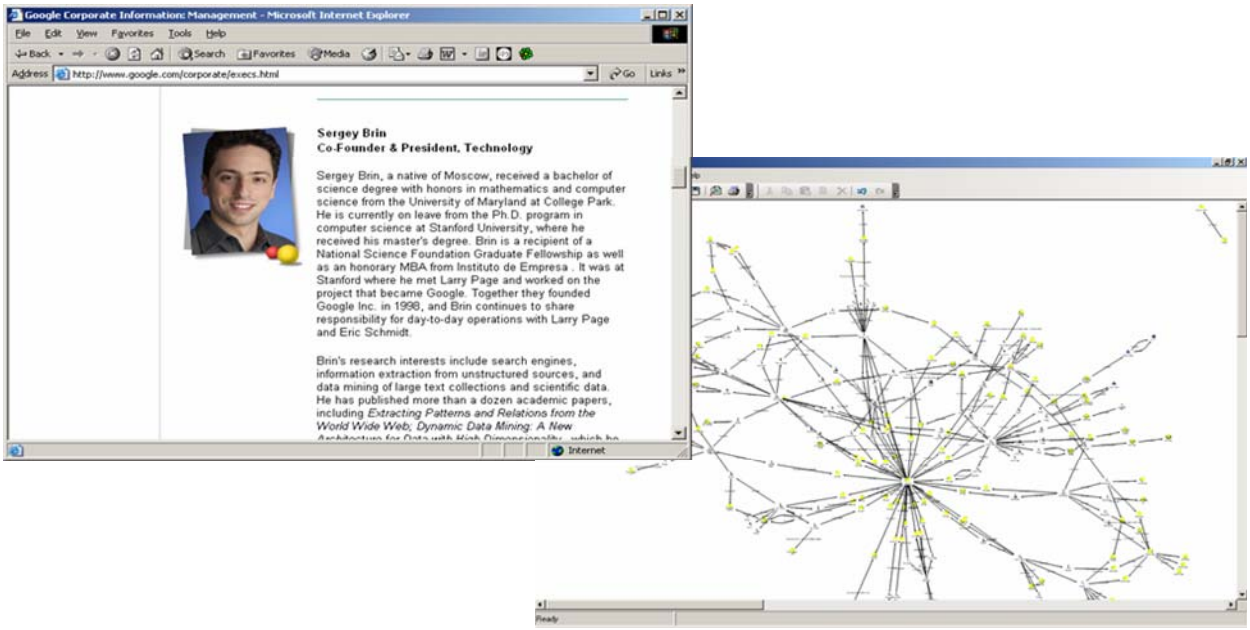

Рис. 13. Обработка страницы Google management Team

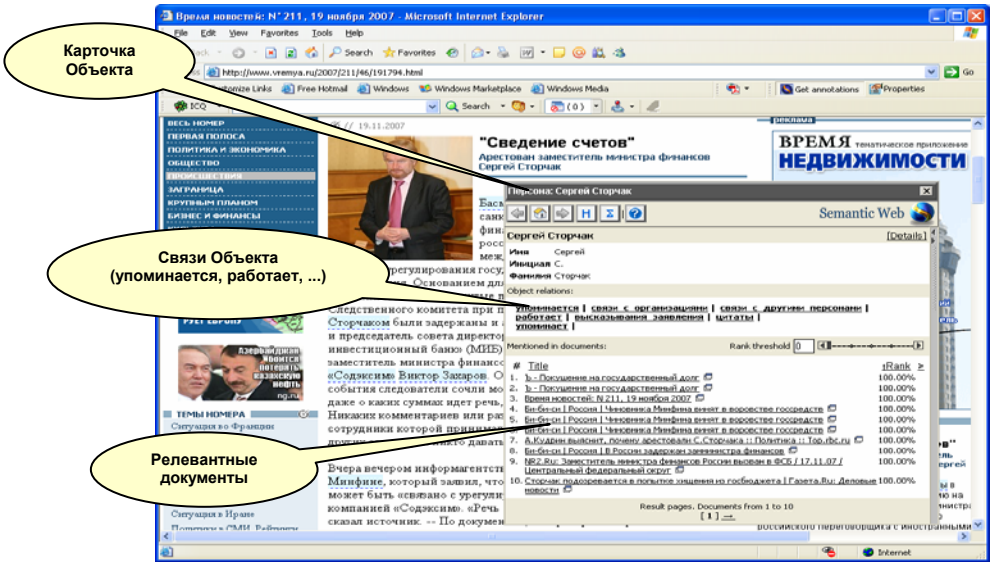

Рис. 14. Экранная форма системы семантической навигации

одного документа или коллекции документов, содержащих информацию, релевантную условиям запроса на его создание. Для каждого фрагмента дайджеста указывается дата публикации документа и гиперссылка на первоисточник.

Сервис семантического реферирования обеспечивает создание текста реферата на основе информации из семантического хранилища знаний с использованием разработанного в компании ЗАО «Авикомп Сервисез» онтологического подхода к реферированию документов [64, 65]. Реферат создается в рамках одного из предопределенных сценариев, каждый из которых определяет сюжет реферата и тот угол зрения, под которым формируется реферат. В настоящее время реализованы два основных сюжета – «пересказ» и «реферат по объекту внимания». В первом случае в процесс генерации текста реферата вовлекаются все объекты и отношения между ними, полученные по запросу пользователя, а во втором – объекты внимания из навигационной карточки.

В 2008-2009 г.г. компания ЗАО «Авикомп Сервисез» выполняла конкурсный проект по

созданию прототипа системы баз данных для внешних пользователей Государственной корпорации «Роснанотех», целью которого было создание системы, автоматизирующей процессы сбора и обработки неструктурированной информации в области наноиндустрии и привлечения экспертных сообществ для разработки понятийной базы знаний о данной предметной области. В общей архитектуре портала Nanotrack основными разделами являются NanoNews, Nanopedia и Analytics.

В первом разделе фиксируются результаты мониторинга заданных информационных источников, а на странице раздела NanoNews, помимо стандартной ленты новостей, обогащенной семантической разметкой, представляются Top-объекты, автоматически полученные в результате обработки, а также интегральная статистическая информация по разделу и/или информация, отфильтрованная в соответствии с запросами пользователя (Рис.15).

Следующим крупным разделом портала Nanotrack.ru является раздел Nanopedia, основная идея которого состоит в автоматической генерации Wiki-страниц на основе результатов семантической обработки информационных материалов. Экранная форма с результатами автоматической генерации Wiki-статьи по теме «углеродные нанотрубки» показана на Рис.16.

Последним крупным разделом портала Nanotrack является раздел Analytics. Базисные функционалы этого раздела связаны с поддержкой принятия решений в следующих основных направлениях: Кто/Что значимо? (Область деятельности, Организации, Персоны, Факты), Как это связано? (Кто делает, Где делает, Когда делает, С кем делает, Что делает) и Как это оценивается? (Что они/о них/об этом говорят, Как они/о них/об этом говорят). Для примера, на Рис. 17 приведена экранная форма страницы портала для сценария «Чем»-«Кто»-«Где» занимается.

В дополнение к указанным функционалам на портале поддерживается геовизуализация результатов.

Оценивая компетенцию ЗАО «Авикомп Сервисез» в сегменте сервисов, при создании которых использовались обработка ЕЯ-текстов и семантические технологии, можно отметить, что здесь накоплен достаточно серьезный опыт разработки и реализации интеллектуальных систем, ядром которых являются средства извлечения информации из текстов на различных естественных языках.

1.4.7. Информационные невидимки

В данном подразделе собрана информация о некоторых проектах и системах, описания которых практически отсутствуют в трудах конференций по тематике настоящего обзора и потому, как правило, трудно доступны для специалистов.

Рис. 15. Экранная форма раздела NanoNews

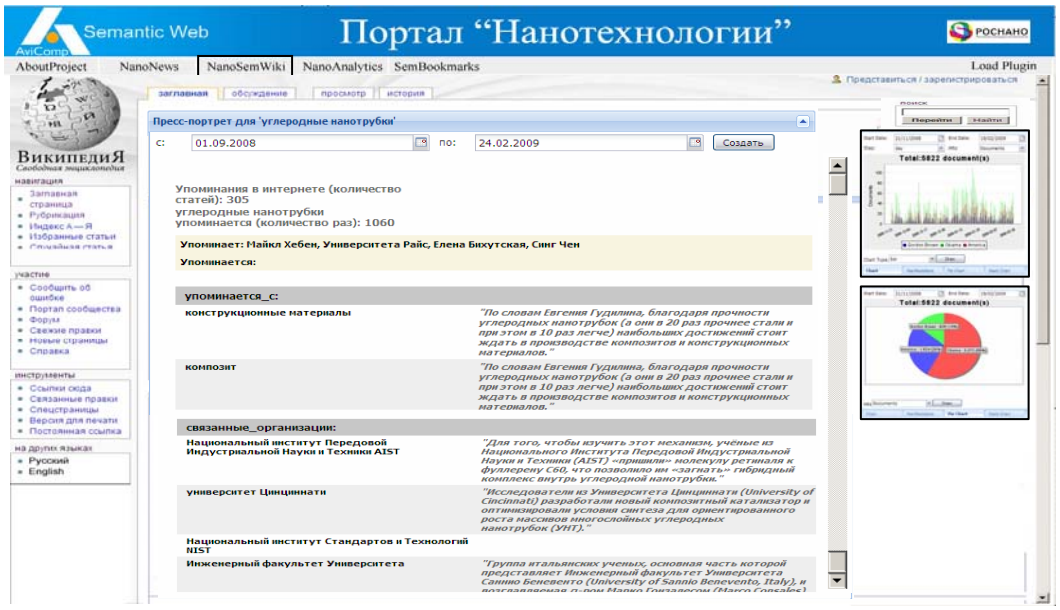

Рис. 16. Wiki-статья по теме «углеродные нанотрубки»

Рис. 17. Экранная форма сценария «Чем»-«Кто»-«Где»

Одним из таких проектов был проект компании «ИнфоСкан» (www.iscann.ru) по созданию информационно-поисковой системы нового типа «КтоТам» (ktotam.ru), которая позволяла извлекать и объединять знания (люди, организации, явления и предметы окружающего нас мира) из массивов открытой публичной информации. Основным ее отличием от множества других поисковых систем было то, что результатом поиска был весь объем знаний, доступный в одном месте, а не множество ссылок на разрозненные сайты. Не менее важным, а порой и более полезным отличием была возможность поиска связей между объектами – людьми, организациями, событиями.

Основу функционала системы составлял поиск людей и организаций. Система «КтоТам» искала людей по фамилии или имени и могла найти людей по дополнительным характеристикам – «Иван хоккеист» и «Сидоров кассир»; позволяла искать людей только по характеристикам, помогая найти всех олимпийских чемпионов из Санкт-Петербурга или всех политиков Краснодара; могла найти связи между людьми, проверяя на практике теорию шести рукопожатий.

Поисковая машина «КтоТам» обрабатывала информацию из тысяч открытых источников (электронные и печатные СМИ, новостные ленты и аналитика, государственные и корпоративные издания, публично доступная информация из социальных сетей и блогов) с учетом лингвистической и логической структуры текста на основе уникальных алгоритмов, позволяющих понимать строение текста, сравнивать и классифицировать информацию. Результатом обработки текста был SmartIndex, представляющий из себя многоуровневый набор фактов и атрибутов о каком либо явлении. Например, для предложения «глава компании Имярек, Александр Краснов, подписал приказ о назначении Ивана Франко руководителем отдела аналитики» SmartIndex может быть специфицирован следующим образом:

глава<id=1. ДОЛЖНОСТЬ. ЗНАЧЕНИЕ-**Глава**>
компании Имярек<id=2. КОМПАНИЯ. НАЗВАНИЕ-**Имярек**>,
Александр Краснов<id=3. ПЕРСОНАЖ. РОД-МУЖСКОЙ. ИМЯ-**Александр**. ФАМИЛИЯ-**Краснов**>,
подписал<id=4. СОБЫТИЕ. ЗНАЧЕНИЕ-**Подписать**. ФОРМА-**Совершенная**>
приказ о назначении<id=5 ФИЗИЧЕСКИЙ ОБЪЕКТ>
Ивана Франко<id=6 ПЕРСОНАЖ. РОД-МУЖСКОЙ. ИМЯ-**Иван**. ФАМИЛИЯ-**Франко**>
руководителем<id=7 ДОЛЖНОСТЬ. ЗНАЧЕНИЕ-**Руководитель**>
отдела аналитики<id=8 КОМПАНИЯ:ПОДРАЗДЕЛЕНИЕ.НАИМЕНОВАНИЕ-**отдел аналитики**.ТИП-отдел>
<СВЯЗЬ id3-id2 ТИП=РАБОТА. ДОЛЖНОСТЬ=id1>
<СВЯЗЬ id6-id8 ТИП=РАБОТА. ДОЛЖНОСТЬ=id7>
.....

Кроме приведенной выше, SmartIndex содержит информацию об источнике фактов, атрибуты, дату и другие дополнительные сведения и строится при каждом упоминании компании Имярек или Александра Краснова и Ивана Франко, что позволяет по конкретному человеку или организации отражать и объединять множество фактов, связей и атрибутов, выделенных из разных источников.

Основное направление усилий команды «КтоТам», по их собственному утверждению, состояло в построении качественных SmartIndex-ов для лингвистически сложных предложений; реализации качественного мета-индекса для всего объема обработанной информации; разработке и реализации алгоритмов объединения информации про одного и того же человека из разных источников и недопущения объединения информации для разных людей (например, однофамильцев или полных тезок).

Оценивая потенциал проекта «КтоТам» в сегменте сервисов, при создании которых использовались средства обработки ЕЯ-текстов, можно констатировать, что у проекта была хорошая идея и интересные решения в области семантического индексирования, но не было серьезного по объему проиндексированного контента, на который могли бы опираться все сервисы. В настоящее время этот проект, по-видимому, закрыт и система «КтоТам» в сети Интернет недоступна.

Компания САЙТЭК (sytech.ru) работает в инновационной сфере информационно-аналитических систем с возможностями интеллектуальной обработки информации. Как показывает анализ доступной информации, компания разрабатывает различные продукты, но с позиций настоящей работы интерес представляет только информационно-аналитическая система «АРИОН». По заявлению разработчиков, это уникальный программный продукт, не имеющий аналогов на российском рынке, и предназначенный для интеллектуальной автоматизированной обработки разнородной информации, использующий инновационные технологии извлечения и обработки знаний. Система позволяет работать как со структурированными (таблицы, базы данных, xml), так и неструктурированными (документы и тексты на естественном языке) источниками информации. Пользователь получает эффективный инструмент аналитики с развитыми механизмами визуализации и большим набором функций по извлечению, загрузке, очистке и обработке информации. На сайте компании есть информация о том, что в основу системы «АРИОН» заложены специализированные алгоритмы обработки информации, разработанные компанией «САЙТЭК» совместно с ИПИ РАН на базе более чем 20-летних теоретических исследований.

Для сбора данных из разнородных источников в ИАС «АРИОН» используется специализированный модуль, для которого могут быть заданы критерии отбора информации из источников и способы ее первичной фильтрации, а также регламенты обновления и загрузки новой информации. Для хранения и обработки доку-

ментов в ИАС «АРИОН» используется специальное полнотекстовое хранилище с набором функций извлечения и загрузки документов из массива, а также автоматической рубрикации документов. В частности, на основе массива документов возможно составление дайджестов и рефератов документов; аннотирование текстов документов; подготовка аналитических подборок; полнотекстовый поиск с учетом морфологии; атрибутивный поиск документов и разметка текстов документов.

Выделение фактографической информации из текстов документов выполняется в лингвистическом процессоре АРИОН-ЛИНГВО, который получает на вход текстовый документ, результатом обработки которого является массив связной фактографической информации. Полученная информация передается в модуль идентификации для выделения похожих и слияния совпадающих объектов. Выделение фактографической информации осуществляется с помощью специализированных правил, которые описывают процедуры выделения объектов и связей на внутрисистемном языке лингвистического процессора, построенном на базе XML.

Набор аналитических функций ИАС «АРИОН» позволяет пользователю эффективно обрабатывать накопленный массив информации, обеспечивая идентификационные возможности, поисково-аналитические функции, а также функции мониторинга и прогнозирования. Информация из ИАС «АРИОН» может быть выведена в виде отчетов в табличном и графическом виде. Предусмотрены режимы агрегирования и подсчета статистики. Результаты статистической обработки представляются в системе в виде таблиц, графиков и гистограмм.

По утверждению разработчиков, ИАС «АРИОН» использует открытые стандарты W3C и имеет возможности взаимодействия со смежными системами. Вся информация в системе «АРИОН» имеет представление в формате xml и может быть выгружена в файл, базу данных, или передана в нужную Вэб-службу.

Программная структура ИАС «АРИОН» базируется на многозвенной архитектуре и позволяет системе встраиваться в существующую вычислительную среду, обеспечивая необходимый уровень производительности и надежности. Доступ пользователей к системе происходит через окно браузера по технологии «тонкого клиента». Для этого предусмотрен информационный портал, обеспечивающий персонификацию данных пользователей и индивидуальные настройки интерфейса.

Если бы все, что перечислено в рекламных материалах по ИАС «АРИОН», было бы реализовано и работало с нужным для практики качеством, то можно было бы согласиться с утверждением разработчиков, что «В настоящее время система не имеет аналогов на российском рынке, как с точки зрения технологичности, так и с точки зрения функциональности и удобства применения». Однако, как показывает анализ материалов и обсуждений данной системы с представителями компании «САЙТЭК», многое из вышесказанного существует либо в виде прототипов, либо находится в разработке. Оценивая потенциал компании «САЙТЭК» в сегменте продуктов на базе обработки ЕЯ, можно констатировать, что здесь более серьезное впечатление оставляют методы обработки данных, а решения в области обработки естественного языка и аналитики на знаниях проработаны слабее. Концепция разработки лингвистических процессоров в целом существенно отличается от других, рассмотренных в настоящей работе, и состоит в том, что лингвисты в данном процессе не нужны, а все необходимые изменения в правилах могут быть выполнены программистами с использованием XML-подобного продукционного языка представления знаний. В компании имеется достаточно мощная инструментальная система поддержки разработки лингвистических процессоров, которая на уровень продуктов не выходит.

Основным направлением программных разработок компании ООО «Аналитические бизнес решения» (www.anbr.ru) является автоматизация деятельности аналитических служб (безопасности, маркетинга, PR-служб, информационно-аналитических отделов и пр.) организаций различных сфер деятельности. Компания с 2004 года специализируется на разработке системы анализа СМИ и конкурентной разведки «Семантический архив», которая, по мнению разработчиков, является одной из самых распространенных систем анализа СМИ и конкурентной разведки в России и странах СНГ. Конкретных данных о методах и качестве обработки ЕЯ-текстов на сайте компании нет, но информация о том, что система «Семантический архив» является современным аналогом систем Кронос, Крос и i2 Analyst Notebook

позволяет позиционировать ее в спектре российских разработок информационно-аналитических систем.

Компания «Ай-Теко», основанная в 1997 году, позиционируется как ведущий российский системный интегратор и поставщик информационных технологий для корпоративных заказчиков, где совершенствуются и активно внедряются собственные информационно-аналитические разработки. С точки зрения настоящего обзора интерес представляют программные продукты компании «Аналитический курьер» и X-Files.

Программный комплекс «Аналитический курьер», как утверждают разработчики, является системой извлечения знаний из документов, где реализованы, помимо прочего:

– параллельная обработка разнородной неструктурированной информации из различных управленческих и юридических документов, сообщений СМИ и информационных агентств, аналитических материалов различного профиля, ресурсов сети Интернет и др.;

– многоязычный семантический поиск с использованием современного тезауруса русского и других языков, обработка запросов на естественном языке для текста на европейских языках;

– выделение сущностей на русском и английском языках, а также определение тональной окраски документов и отдельных объектов, включая выделение упоминаний и цитирования;

– определение индекса информационной значимости объектов мониторинга;

– выявление ключевых тем документа, коллекции документов и построение их взаимосвязей в виде семантической сети;

– автоматическое общее и тематическое реферирование коллекций или отдельных документов, построение дайджестов по каждому объекту или теме документа и регламентный выпуск аналитических отчетов.

Реализована система на Windows-платформе NET, имеет трехзвенную архитектуру с «тонким» клиентом и предоставляет пользователям Web-интерфейс. Для хранилища аналитических данных испльзуются СУБД MS SQL Server и ORACLE.

Полученную по результатам поискового запроса подборку сначала обрабатывает семантический процессор, выделяющий сущности из документов. Затем аналитический процессор по документам и сущностям формирует решетку формальных понятий, на основании анализа которой определяются и удаляются похожие документы, а также повторяющиеся или незначимые сущности, в результате чего остаются только документы и сущности, независимые друг от друга. На основании общих значимых сущностей документы исходной подборки разделяются на кластеры, для визуализации которых строится многоуровневое дерево.

Важнейшим результатом компании, по утверждению разработчиков системы «Аналитический курьер», является развитие компонентов лингво-семантического анализа текста на русском и английском языках. При этом в рамках лингвистического анализа текстов реализован лексический и морфологический анализ текста, его предсинтаксический и синтаксический анализ (построение дерева разбора предложения и определение синтаксических ролей слов в предложении: подлежащее, сказуемое, дополнение, обстоятельство и т.д.), а также постсинтаксический анализ (выделение типизированных сущностей). Последующий семантический анализ осуществляет типизацию сущностей (физические, юридические лица; одушевленные предметы; даты; регионы и др.), и их нормализацию. Для идентификации ссылочно представленных сущностей используются эвристические методы. Выделение таких типов сущностей, как адреса, телефоны и т.п., производится с помощью расширяемых (в том числе и пользователем) правил.

Как утверждается в пресс-релизе компании по системе «Аналитический курьер», компанией завершена разработка нового современного тезауруса русского языка, совместимого со стандартом WordNet 3.0, в составе которого более 160 тысяч групп синонимов, 700 тысяч связей между ними, 170 тысяч лексем и 13 типов семантических отношений. Для управления тезаурусом разработан Web-сервис, который может быть использован как в системах «Аналитический курьер» и X-Files, так и в других системах.

Система управления досье X-Files предназначена для решения задач выделения достоверных фактов из различных источников, заполнения ими досье на объекты мониторинга и их последующей аналитической обработки. Она используется для обеспечения процессов принятия решений при наличии большого объ-

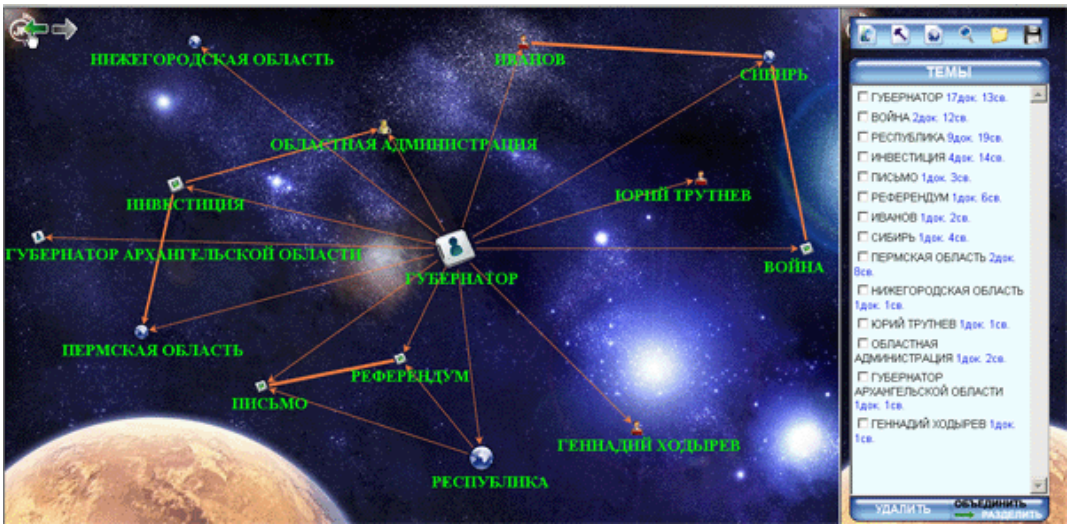

Рис. 18. Пример семантической карты по аспектам деятельности губернаторов

ема «сырого» контента, что характерно для деятельности органов государственной власти, правоохранительных органов, крупных коммерческих компаний. Система позволяет извлекать факты из Интернета, корпоративных источников документов, из учетных баз данных и других информационных источников, в частности, из хранилища документов системы «Аналитический курьер». Для выделения из текста сущностей, отношений между ними и других свойств фактов система X-Files , по-видимому, использует лингвистический процессор системы «Аналитический курьер», а также общий и предметные тезаурусы.

Как утверждают разработчики, X-Files является единственной на российском рынке системой, реализующей максимальную полноту и точность поиска фактов, единственной на российском рынке системой, формирующей хранилище фактов, что обеспечивает возможность выработки и хранения гипотез о вероятных связях объектов при отсутствии фактов об этих связях путем порождения соответствующих гипотез либо при наличии у каждого из пары объектов устойчивых связей с третьими общими для них объектами, либо при наличии для каждого из объектов фактов с общими местами совершения фактов в общем интервале времени. При этом поддерживается высокопроизводительная аналитическая обработка большого числа фактов (более 100 000), поиск кратчайших и эффективных связей объектов с включением связей-гипотез и анализ динамики фактов, хранение фактов, полученных как в источниках учетной, текстовой, так и видео- информации, а также совместная аналитическая обработка фактов, представленных на разных языках, ассоциативный поиск фактов за счет самообучения правил идентификации объектов и идентификация тонально окрашенных фактов с последующим включением их в досье объектов.

Таким образом, система X-Files структурирует смысл документов, автоматически выявляя факты нужного типа, связанные с объектами, на которых системой автоматически ведется досье, а выявленные структурированные факты позволяют находить скрытые причины событий или спрогнозировать поведение объектов в будущем. При этом взаимосвязи объектов визуально представляются в виде карты связей, один из примеров которых представлен на Рис. 18, заимствованного с сайта компании «Ай-Теко».

Приведенный выше пример, как и другие иллюстрации функционирования систем «Аналитический курьер» и X-Files, показывают, что в данном случае семантически значимые именованные связи между объектами не идентифицируются, а отсутствие данных о точности и полноте выделения этими системами именованных сущностей не позволяет с достоверностью оценить качество системы лингвосемантического анализа.

Вместе с тем, как представляется из анализа представленной на сайте компании информации, в рассмотренных выше системах реализованы достаточно мощные и интересные методы индексации текстов, кластеризации коллекций

документов и неплохие средства визуализации полученных результатов. Серьезное впечатление оставляет и тезаурус русского языка, который, по нашему мнению, может с успехом использоваться для семантической индексации ЕЯ-текстов.

Заключение

Обсуждение представленных в [1, 2] и в настоящей работе результатов в области извлечения информации из текстов, формирования пространств знаний и в области Semantic Web показывает, что соответствующие исследования и разработки активно ведутся во всем мире. При этом лидирующее положение занимают исследовательские коллективы и компании США, в которых имеются мощные R&D подразделения с многолетним опытом работ в данной наукоемкой области. Результаты европейских стран в целом отстают от результатов США, хотя отдельные исследовательские коллективы находятся на мировом уровне и активно участвуют в международных проектах по данной проблематике. В последнее время в указанных направлениях наблюдается серьезная научно-техническая и, особенно, патентная активность стран Юго-Восточной Азии.

Российские исследования и разработки в целом несколько отстают от европейских и американских. Вместе с тем, и среди российских исследовательских коллективов и компаний, способных создавать наукоемкую продукцию, имеются примеры разработок европейского и даже мирового уровня. Следует также отметить, что для российских исследований и разработок характерным является высокий научный уровень проработки отдельных аспектов создания методов автоматической обработки естественного языка и, на этом фоне, недостаточно высокий уровень реализации своих идей. С организационной точки зрения следует отметить тенденцию объединения исследовательских коллективов из России с коллективами российских и иностранных компаний, работающих в данном сегменте рынка. При этом, как правило, соответствующая компания интегрирует высококвалифицированную команду российских исследователей в свои разработки, что может, спустя некоторое время, привести к снижению теоретического уровня работ и, как следствие, увеличению разрыва между исследованиями и разработками российских ученых и специалистов и зарубежными разработками.

Проведенный в настоящей работе анализ российских исследований и разработок в области извлечения информации из текстов на естественных языках показывает, что в исследовательских коллективах, научных организациях и коммерческих компаниях ведутся активные работы в этом направлении, оценивая которые можно констатировать:

– российские исследования и разработки достаточно хорошо коррелируют с общемировыми тенденциями в этой области;

– результаты работ начинают использоваться на практике;

– многие исследования концентрируются на отдельных аспектах создания систем извлечения информации из текстов и лишь отдельные разработки имеют комплексный характер и доводятся до реально действующих систем.

Литература

1. Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1). Искусственный интеллект и принятие решений. № 1.2008.
2. Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 2). Искусственный интеллект и принятие решений. № 4.2009.
3. Хорошевский В.Ф. Извлечение информации из текстов на конференциях серии ДИАЛОГ: взгляд соседа по лестничной клетке. // Труды международной конференции "Диалог 2010". М.: Наука. 2010.
4. Сайт инициативы РОМИП. http://romip.ru/about.html.
5. Kononenko I., Kononenko S., Popov I., Zagorul'ko Yu. Information Extraction from Non-Segmented Text (on the material of weather forecast telegrams). // Content-Based Multimedia Information Access. RIAO'2000 Conference Proceedings, v.2. 2000.
6. Сидорова Е.А. Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008.
7. Загоруйко Н.Г., Налетов А.М. Гребенкин И.М. На пути к автоматическому построению онтологии. // Труды конференции "Диалог-2003".
8. Кузнецов И.П. Семантические представления. // М.: Наука. 1986.
9. Кузнецов В.П., Мацкевич А.Г. Автоматическое выявление из документов значимой информации с помощью шаблонных слов и контекста. // Труды международного семинара "Диалог-1998" по компьютерной лингвистике и ее приложениям. Том 2. Казань. 1998.
10. Ермаков А.Е. Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Труды международной конференции

"Диалог- 2007" «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука. 2007.
11. Добров Б.В., Лукашевич Н.В. Онтологии для автоматической обработки текстов: Описание понятий и лексических значений // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог -2006". Бекасово. 31 мая – 4 июня 2006 г. 2006.
12. Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е. Методы и средства автоматизированного проектирования прикладной онтологии // Известия РАН. Теория и системы управления. № 6.М.: 2004.
13. Ермаков А.Е., Автоматизация онтологического инжиниринга в системах извлечения знаний из текста, // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции "Диалог-2008". – М.: Наука. 2008.
14. Антонов А., Курзинер Е. Автоматическое выделение предметной области большого необработанного текстового массива // Компьютерная лингвистика и интеллектуальные технологии, Труды Международного семинара "Диалог-2002".
15. Баглей С.Г., Антонов А.В., Мешков В.С., Суханов А.В. Кластеризация документов с использованием метаинформации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006.
16. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007.
17. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. М.: Физматлит. 2006.
18. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции». 2003.
19. Лукашевич Н.В., Добров Б.В. Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара "Диалог-2002". – М.: Наука. 2002.
20. Браславский П.И., Соколов Е.А. Автоматическое извлечение терминологии с использованием поисковых машин интернета // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007.
21. Браславский П.И., Соколов Е.А. Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008.
22. Хорошевский В.Ф., Проект OntosMiner: воспоминания о будущем. //Труды конференции КИИ-2010, Тверь. 2010.
23. Cunningham H., Maynard D., Bontcheva K., Tablan V. GATE: an Architecture for Development of Robust HLT Applications. // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002.
24. Berners-Lee T. The Semantic Web and Research Challenges, http://www.w3.org/2003/Talks/01-sweb-tbl/slide1-0.html.
25. Efimenko I., Minor S., Starostin A., Drobyazko G., Khoroshevsky V. Generating Semantic Content for the Next Generation Web, Chapter in Monograph "Semantic Web", Publ. IN-TECH, 2009, ISBN 978-953-7619-33-6.
26. Хорошевский В.Ф. OntosMiner: семейство систем извлечения информации из мультиязычных коллекций документов, Труды конференции КИИ-2004. Тверь. 2004.
27. Справочник "Искусственный интеллект". Кн. 3. "Программные и аппаратные средства» (Захаров В.Н., Хорошевский В.Ф. – ред.).М.: Сов.Радио. 1990.
28. V. Karasev, O. Mishchenko, A. Shafirin, Interactive Debugger of Linguistic Programs in the GATE Environment, In: Proc. International Workshop "Information Extraction for Slavonic and Other Central and Eastern European Languages". IESL-2003. Borovec. Bulgaria. 2003.
29. Karasev V., Khoroshevsky V.F., Shafirin A., New Flexible KRL JAPE+: Development & Implementation, In Proc. of Joint Conference on Knowledge-Based Software Engineering 2004, JCKBSE-2004, 24-27 August 2004, Protvino.
30. Appelt D. The Common Pattern Specification Language. Technical report, SRI International, Artificial Intelligence Center. 1996.
31. Мальковский М.Г., Старостин А.С. Система Treeton: анализ под управлением штрафной функции, Программные продукты и системы. № 1. 2009.
32. Кузнецов И.П. Механизмы обработки семантической информации. М.: Наука. 1978.
33. Кузнецов И.П., Шарнин М.М. Интеллектуальный редактор знаний на основе расширенных семантических сетей. // Сборник «Системы и средства информатики», Вып. 5., М.: Наука. 1993.
34. Клоксин У., Меллиш К. Программирование на языке ПРОЛОГ. –М.: Мир. 1987.
35. Hewitt C. The repeated demise of logic programming and why it will be reincarnated What Went Wrong and Why: Lessons from AI Research and Applications. Technical Report SS-06-08. AAAI Press. March 2006.
36. Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor : компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов. М.: 2003.
37. Thakker D., Osman T., Lakin P. GATE JAPE Grammar Tutorial, February 27, 2009. http://gate.ac.uk/sale/thakker-jape-tutorial/GATE JAPE manual.pdf
38. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. — М.: Прогресс, 1981. — Вып. 10.
39. Мельчук И.А. Опыт теории лингвистических моделей "Смысл - Текст" // М.: Наука.1974.
40. Апресян Ю.Д. Избранные труды, том II. Интегральное описание языка и системная лексикография. // М. 1995.
41. Попов Э.В. Общение с ЭВМ на естественном языке. – М.: Наука. 1982.

42. Дракин В.И., Попов Э.В., Преображенский А.Б. Общение конечных пользователей с системами обработки данных. – М.: Радио и связь. 1988.
43. Справочник "Искусственный интеллект". Кн. 1, "Системы общения и экспертные системы» (Попов Э. В. – ред.).М.: Сов.Радио. 1990.
44. Большакова Е.И., Баева Н.В. Автоматический анализ дискурсивной структуры научного текста. Труды Международной конференции “Диалог-2004” / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П.Селегея – М.: Наука. 2004.
45. Рябцева Н.К. Ментальные перформативы в научном дискурсе // Вопросы языкознания. № 4.1992.
46. Вежбицка А. Метатекст в тексте // Новое в зарубежной лингвистике. Вып. VIII . М.: Прогресс. 1978.
47. Севбо И.П. Сквозной анализ как шаг к структурированию текста // НТИ. Сер. 2. № 2.1998.
48. Баева Н.В., Большакова Е.И., Васильева Н.Э. Структурирование и извлечение знаний, представленных в научных текстах // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х томах. Т. 2. – М.: Физматлит. 2004.
49. Большакова Е.И. О принципах построения компьютерного словаря общенаучной лексики //Труды Международного семинара “Диалог -2002” по комп. лингвистике и интеллект. технологиям.Т.1. М. 2002.
50. Bolshakova, E.I. Phraseological Database Extended by Educational Material for Learning Scientific Style. In: ACH/ALLC 2001: The 2001 Joint International Conference. New York University, New York, 2001.
51. Старостин А.С., Арефьев Н.В., Мальковский М.Г. Синтаксический анализатор «». Принцип динамического ранжирования гипотез, // Труды международной конференции "Диалог- 2010". М.: Наука . 2010.
52. Orasan C., Cristea D., Mitkov R. and Branco A. Anaphora Resolution Exercise – an overview. Proceedings of 6th Language Resources and Evaluation Conference (LREC2008), Marrakesh, Morocco, 28 – 30 May 2008.
53. Grishman R. TIPSTER Text Architecture Design. Version 3.1 7 October 1998, New York University, 1998.
54. Кузнецов И.П., Ефимов Д.А. Особенности извлечения знаний семантико-ориентированным лингвистическим процессором Semantix.// Сб. Компьютерная лингвистика и интеллектуальные технологии. Вып. 7 (14). По материалам конференции «Диалог- 2008». М.: РГГУ.2008.
55. Кузнецов И.П., Ефимов Д.А. Средства настройки процессора semantix на предметную область, Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27-31 мая 2009 г.). Вып. 8 (15). М. : РГГУ. 2009.
56. Кузнецов И.П., Сомин Н.В. Средства настройки семантико-ориентированного лингвистического процессора на выделение и поиск объектов. Сб. ИПИ РАН. Вып.18. 2008.
57. Kuznetsov I.P., Kozerenko E.B. Linguistic Processor “Semantix” for Knowledge extraction from natural texts in Russia and English. Proceeding of International Conference on Machine Learning, ISAT-2008. 14-18 July, 2008 Las Vegas, USA// CSREA Press. 2008.
58. Ермаков А. Е. Извлечение знаний из текста и их обработка: состояние и перспективы. Информационные технологии. №7. 2009.
59. Engels R., Bremdal B. Information Extraction: State-of-the-Art Report, CognIT a.s., Asker, Norway, 2000.
60. Тихомиров И. А., Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus. //Труды четвертого российского семинара по оценке методов информационного поиска РОМИП'2006. Санкт-Петербург: НУ ЦСИ. 2006.
61. Осипов Г.С., Тихомиров И.А., Смирнов И.В., Семантический поиск в сети интернет средствами поисковой машины Exactus, //Труды 11-й национальной конференции по искусственному интеллекту с международным участием КИИ-2008. М.: Физматлит. 2008.
62. Золотова Г.А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. Институт русского языка РАН им. В. В. Виноградова. М. 2004. – 544 с.
63. Осипов Г.С. Приобретение знаний интеллектуальными системами. Основы теории и технологии - М.: НАУКА ФИЗМАТЛИТ. 1997.
64. Ефименко И.В. Лингвистические аспекты кросс-языкового реферирования: синтез текстов под управлением предметных онтологий, В Сб. Трудов 10-й Конференции по искусственному интеллекту, КИИ-2006. Обнинск. 2006.
65. Кананыкина П.Г., Хорошевский В.Ф. Интеллектуальное реферирование: онтологический подход и его реализация в решениях Ontos. В Сб. Трудов 11-й Конференции по искусственному интеллекту, КИИ-2008. Дубна. 2008.

Хорошевский Владимир Федорович. Заведующий сектором Вычислительного центра РАН им. А.А. Дородницына. В 1971 году окончил Московский инженерно-физический институт, доктор технических наук, профессор. Опубликовал более 100 печатных работ, среди которых 4 монографии и 5 учебных пособий. Область научных интересов: программное обеспечение систем искусственного интеллекта, представление знаний, обработка естественного языка, мультиагентные системы, семантические технологии, семантический Веб. E-mail: khor@ccas.ru.