В.Ф. Хорошевский

Пространства знаний в сети Интернет и Semantic Web (Часть 1)

Аннотация. Представлено новое направление на стыке работ по искусственному интеллекту и Интернет-технологиям – Семантический Вэб. Дается обзор состояния исследований, рассматриваются проблемы организации пространств знаний в Интернет, методы и средства извлечения знаний из текстов на естественных языках, а также вопросы использования пространств знаний при создании прикладных интеллектуальных систем, функционирующих в Интернет.

Ключевые слова: извлечение знаний, интеллектуальные системы, семантический Вэб.

Введение

Исследования IDC [1], мирового лидера в прогнозных исследованиях показывают, что к 2008 году количество информации, хранящейся в компьютерных системах, составит 5444 петабайт, притом, что в 2003 году оно было около 831 петабайт, а все, опубликованное человечеством в книгах, составило порядка 200 петабайт. Как отмечается в работе [2], "...до 2020 года количество информации и потребности в ней будут расти экспоненциально... Без умения создавать и обрабатывать такие объемы информации ЛПР будущего будут введены в состояние, которое можно назвать "аналитический паралич"...". Таким образом, одной из самых больших проблем современного общества является информационное переполнение, которое, в значительной мере, определяется сетью Интернет – всемирного хранилища, «открывающего» доступ к миллионам и миллиардам различных информационных ресурсов, независимо от их географической и национальной локализации. Однако поиск и использование нужной информации становится все более сложным, трудоемким и неэффективным, несмотря на огромные усилия (как научно-технические, так и организационно-финансовые) по увеличению эффективности доступа и обработки уже существующей и постоянно появляющейся новой информации. И, по сути дела, в настоящее время мировым сообществом уже осознано направление главного «удара» в борьбе с информационным взрывом – переход от хранения и обработки данных к накоплению и обработке знаний. А один из подходов, в рамках которого для решения вышеуказанной проблемы сейчас сосредоточиваются значительные научно-технические ресурсы, - переход от классического Интернет (WWW) к семантическому (Semantic Web).

Учитывая вышесказанное, целью настоящей статьи является аналитический обзор состояния исследований, связанных с формированием и использованием пространств знаний в сети Интернет – направления, активно развивающегося на стыке работ в области искусственного интеллекта, компьютерной лингвистики и Интернет-технологий. Организовано изложение следующим образом: сначала приводится краткая история вопроса и рассматриваются основные проблемы организации пространств знаний в сети Интернет и возникающие при этом задачи, а затем, более подробно, обсуждается проблема извлечения знаний из текстов на естественных языках, а также методы и средства представления и манипулирования знаниями, извлеченными из текстов. В заключительной части работы представлены перспективные направления использования пространств знаний в интеллектуальных Интернет-системах.

В силу большого объема статья состоит из двух частей. В первой части основной объем материала связан с обсуждением решений, продуктов и систем, разрабатываемых в зарубежных коллективах и организациях. В следующей части данной работы предполагается провести обзор решений, продуктов и систем, разрабатываемых в России и странах СНГ, а также остановиться более подробно на вопросах семантической навигации по пространствам знаний и прикладных интеллектуальных системах, ориентированных на Семантический Вэб.

1. Semantic Web: краткая история вопроса и основные проблемы

Концепция Semantic Web (SW), которую на международной конференции XML-2000 выдвинул Тим Бернерс-Ли (Tim Berners-Lee) – один из основоположников WWW и нынешний председатель WWW-консорциума (W3C), заключается в организации такого представления информации в сети, чтобы допускалась не только ее визуализация, как это происходит сейчас, но и эффективная автоматическая обработка [3]. По определению W3C, Semantic Web представляет собой расширение WWW, в рамках которого информация (Web-контент) представляется в форматах, обеспечивающих ее использование программными агентами, позволяя им, таким образом, искать, разделять и интегрировать информацию значительно легче, чем это происходит сейчас.

С момента появления SW-концепции прошло уже более 5 лет и сейчас специалисты говорят о грядущей семантической волне (Semantic Wave), которая, по оценке руководителя проекта Project 10X Миллса Дэвиса [4], существенным образом изменит характер работы с информацией. В последнем. аналитическом отчете Gartner Group [5] дается прогноз, что к 2012 году в 80% общедоступных веб-сайтов будет, в той или иной степени, использоваться семантический гипертекст для создания семантических веб-документов (с вероятностью 0.7), а к 2012 году в 15% общедоступных веб-сайтов будут использоваться развитые веб-онтологии для создания семантических баз данных (с вероятностью 0.6).

Вместе с тем, следует отметить, что SW-эра, в отличие от эпохи Интернет, еще только приближается и на этом пути существует значительное число научных, технических, технологических и чисто человеческих проблем, основными из которых являются [6]:

- доступность семантического контента;
- доступность онтологий и средств их разработки; а также эволюция онтологий;
- масштабируемость;
- мультиязыковость;
- визуализация и
- стабильность.

Доступность семантического контента является основной проблемой на пути формирования и использования пространств знаний, так как основная масса информации на Web не представлена в SW-форматах и нет надежды, что эта работа может быть выполнена вручную.

Онтологии, по мнению практически всех специалистов, являются ключевым компонентом в решении проблемы семантизации Web-контента. В связи с этим особое значение приобретают проблемы онтологического инжиниринга (методы и средства разработки и эволюции онтологий), а также доступность уже существующих онтологий.

Значительные усилия должны быть предприняты для хранения, обработки и поиска семантического контента, причем решения в этой области должны обеспечивать эффективную работу с огромными объемами знаний.

Проблема мультиязыковости контента существует и в классическом Web, но для SW, который по своей сути должен поддерживать эффективный доступ к информации независимо от того, на каком языке она представлена изначально, эта проблема является одной из основных. И решение ее специалисты, в значительной мере, связывают с решением проблем онтологического инжиниринга.

Представление информации для пользователей (визуализация контента) также должно претерпеть существенные изменения и обеспечить свободную ориентацию в огромном количестве фактов, которые отвечают его потребностям.

Последняя по счету, но не по важности, проблема связана с обеспечением стабильность SW, а это, в свою очередь, предполагает, что безотлагательные усилия должны быть предприняты в области стандартизации, обеспечивающей создание технологий, необходимых для формирования пространств знаний.

Понятно, что подробное рассмотрение и обсуждение всех перечисленных выше проблем в рамках одной статьи является практически безнадежным делом. Поэтому ниже мы, в основном, сосредоточимся на формировании семантического контента, его визуализации и навигации по пространствам знаний, затрагивая остальные проблемы SW лишь в связи с вышеуказанными и предполагая, что они еще ждут своих авторов соответствующих аналитических обзоров.

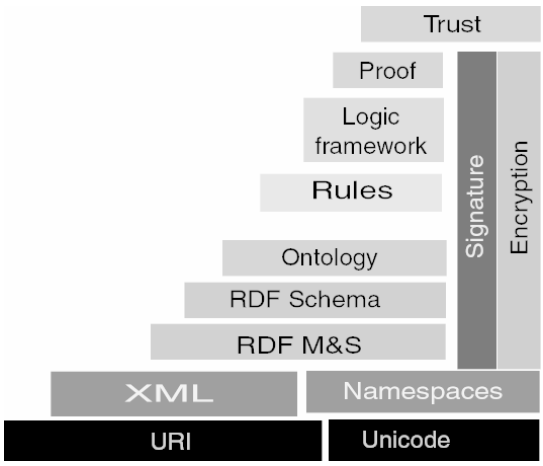

Рис. 1. «Слоеный пирог» Semantic Web

2. Формирование пространств знаний в сети Интернет

2.1. Предварительные замечания

По нашему мнению, обозначенная выше проблема стандартизации является одной из ключевых проблем SW и с этой точки зрения представляет интерес хотя бы краткое обсуждение схемы, представленной на Рис. 1, которую специалисты называют «слоеным пирогом» Тима Бернерс-Ли [7].

Для нижних уровней этой схемы, которые можно объединить в слой «RDF-данные», общая цель, сформулированная W3C, была следующей: разработка форматов сериализации данных и интероперабельность приложений. В результате усилий исследовательских коллективов и консорциума W3C для этого слоя были разработаны и реализованы рекомендации по форматам XML, Namespace (пространства имен) и RDF, которые в настоящее время существуют на уровне стандартов de facto. И можно констатировать, что результаты по данному направлению уже перешли из стадии исследований в стадию использования, в том числе и в коммерческих системах. На уровне RDF-схем предложены и поддерживаются W3C стандарты RDFS (RDF-схем), которые позволяют специфицировать словари используемых терминов, и разрабатываются соответствующие спецификации для существующих и новых приложений. Здесь, как и в предыдущем направлении, результаты уже перешли из стадии исследований в стадию использования.

На онтологическом уровне (Ontology) «слоеного пирога» SW ситуация несколько иная. По сути дела, в этом направлении был достаточно мощный задел в рамках исследований по представлению знаний (в частности, общие подходы к представлению знаний типа фреймов и семантических сетей, конкретные формализмы, языки и системы представления знаний (например, Frame Logics [8], SHOE [9] и др). Вместе с тем, работа по стандартизации средств представления знаний онтологического уровня далеко не закончена, а создание соответствующих средств онтологического инжиниринга является в настоящее время одной из «горячих точек» в данной области. Основными направлениями исследований и разработок здесь являются создание более мощных средств спецификации онтологий, обеспечивающих вывод на знаниях и проверку целостности знаний, средств поддержки целостности онтологических спецификаций в процессе эволюции как спецификаций самих моделей, так и стандартов, а также средств спецификации перекрестных ссылок между словарями и конвертирования спецификаций. Основным результатом в этом направлении можно считать «выравнивание» средств спецификации онтологий, разработанных в США (DAML) и в Европе (OIL), до общего формализма (DAML+OIL [10]), а также разработку консорциумом W3C стандарта de facto на спецификацию онтологий – языка OWL [11].

Понятно, что с переходом к верхним слоям схемы, представленной на Рис. 1, ситуация смещается от уже разработанных стандартов и их использования к исследованиям в соответствующих направлениях. И в этом смысле показателен промежуточный слой – «Слой правил»

(Rules), проработка которого потребовала внесения изменений в слои предыдущих уровней (так, например, были добавлены средства спецификации переменных в RDF-слои), а также поиска новых выразительных и простых средств спецификации отношений и средств для спецификации запросов к базам знаний с возможностью фильтрации получаемых результатов, аналогичных SQL. Наряду с этим в данном направлении ведутся исследования и разработки по теории монотонных и немонотонных систем вывода на правилах, а также работы по созданию новых приложений с использованием технологий типа «webized rule engine technology». Существующую ситуацию в этом направлении можно охарактеризовать следующим образом: уже существуют разные системы спецификации правил и требуется их сравнительный анализ, «вэбизация» и стандартизация, разработан язык SPARQL, который постепенно становится стандартом de facto на язык запросов к базам знаний.

На уровнях «Логические Основы» (Logic Framework) и «Подтверждение» (Proof) ситуация еще больше смещается в область фундаментальных исследований. Здесь предлагаются подходы к спецификации аксиом для систем, основанных на правилах, исследуются различные логики, причем основное внимание уделяется системам, в которых не выполняется аксиома «замкнутого мира», предлагаются средства валидации доказательств. Но пока нет основы для стандартизации систем, основанных на правилах, а существующие системы правил легко экспортируются из разных систем, но плохо импортируются в другие системы.

Что касается уровня «Доверие» (Trust), то здесь только формируются направления фундаментальных и прикладных исследований, поскольку все утверждения в Web-среде существуют в некотором контексте и приложения должны учитывать эти контексты, так как нельзя считать, что все факты, полученные из сети, являются истинными. Таким образом, в данном направлении существует широкое поле для исследований и последующей стандартизации.

Однако, оценивая ситуацию по стандартизации в рамках SW в целом, можно констатировать, что уже создан базис стандартизации в виде XML-, RDF(S)-, OWL- и SPARQL-спецификаций, на которые могут опираться работы по семантизации контента и его использованию в различных приложениях.

Как отмечалось выше, одной из основных целей создания SW является «семантизация» контента, существующего в настоящее время в рамках классического WWW, и вновь создаваемого контента, что должно обеспечить интеллектуальное его использование программными агентами для решения задач, значимых для пользователей. По сути дела, это требует семантического аннотирования контента на базе соответствующих онтологий, где фиксируется смысл отдельных его элементов и связей между ними.

Работы данного направления имеют достаточно представительную историю и начинались в сообществе специалистов по искусственному интеллекту [12,13,14,15], где были заложены теоретические основы представления и манипулирования знаниями в Интернет и разработаны прототипы инструментальных средств семантического аннотирования. В дальнейшем центр тяжести этих исследований и разработок переместился в Интернет-сообщество с ориентацией на SW [16]. В настоящее время основные усилия специалистов сосредоточены на создании методов и средств автоматического и/или автоматизированного аннотирования контента под управлением онтологий, причем в рассмотрение вовлекается не только статический контент, представленный в Интернет, но и информация из баз данных и других источников (так называемый Deep Web, объем которого, по некоторым оценкам, на порядок больше).

К решению проблемы семантизации контента существует несколько подходов. В рамках одного из них, активно поддерживаемого консорциумом W3C, для семантического аннотирования предлагается использовать RDF(S) [17] и/или OWL [18]. Подход RDF/OWL, по мнению Алекса Искольда (Alex Iskold), является мощным и перспективным, но достаточно сложным для понимания и использования основной массой специалистов, занятых аннотированием. Более того, такой подход предполагает наличие мощных инструментальных средств конвертирования существующего HTML-контента в RDF/OWL метаданные, обеспечивающих, как минимум, 80% автоматизацию, оставляя 20% на диалог с пользователем для завершения всей работы. Другой известный подход – Микро-

форматы (Microformats) [19], который появился в связи с осознанием сложности RDF/OWL-подхода, позволяет «имплантировать» семантические тэги в существующие HTML-страницы с помощью достаточно простых инструментальных средств на основе заранее определенных спецификаций их семантики. В настоящее время многие популярные сайты (например, Facebook, Yahoo! Local) используют этот подход для аннотирования событий на своих HTML-страницах.

Вместе с тем, следует отметить, что и в том, и в другом случае ядром соответствующих инструментальных средств могут и должны стать системы понимания естественного языка (Natural Language Understanding).

2.2. Извлечение информации из текстов на естественных языках

Автоматическая обработка естественноязыковых текстов всегда была в фокусе интересов компьютерной лингвистики и искусственного интеллекта, а лет 5-10 назад в этой области начался новый всплеск работ, причем не только в исследовательских коллективах, но и в сфере индустрии информационных технологий (ИТ). Особую важность такие работы имеют для SW, так как в системах семантического аннотирования документов «узким горлышком» является автоматическая обработка естественного языка (ЕЯ).

2.2.1. Исторические замечания

Ретроспективный анализ исследований и разработок по компьютерной обработке ЕЯ показывает [20], что развитие данной области уже насчитывает несколько периодов:

- *60-е годы – середина 70-х годов XX столетия.* Разработка формальных моделей и методов, накопление начального опыта в прототипизации ЕЯ-систем.
- *Середина 70-х годов – 80-е годы XX столетия.* Разработка методов и средств обработки ЕЯ, создание первых промышленных систем общения с базами данных на ЕЯ.
- *Середина 80-х годов – середина 90-х годов XX столетия.* Разработка когнитивных моделей понимания ЕЯ и прототипов систем, использующих модели мира для понимания языка.
- *Середина 90-х годов XX столетия – начало XXI века.* Переход от лингвистики приложения к лингвистике текста, разработка методов и средств обработки ЕЯ-текстов. Появление первых коммерческих систем обработки ЕЯ-текстов.

К основным результатам первых четырех периодов можно отнести то, что

- были выделены классы ЕЯ-систем и их функциональные компоненты, причем состав и основные функции выделенных компонент, их разбиение на подсистемы остается устойчивым и на современном этапе;

- в области моделей и методов анализа ЕЯ получены теоретически и практически значимые морфологические модели анализа\синтеза лексем; разработаны модели синтаксического анализа основных ЕЯ-конструкций; предложен спектр методов реализации основных моделей анализа ЕЯ-конструкций, которые могут использоваться на практике; выделены основные приемы эвристической реализации частных моделей интерпретации ЕЯ-высказываний; проработаны частные модели концептуального синтеза ЕЯ-текстов; предложены и практически проверены модели и методы лингвистического синтеза;

- в области моделей понимания разработаны многоуровневые модели, учитывающие не только лингвистические, но и когнитивные составляющие этого процесса;

- в области реализации разработаны прототипы интеллектуальных ЕЯ-систем; имеются промышленные реализации ЕЯ-систем разного класса, которые, в большинстве случаев, лишь «имитируют» полномасштабное понимание ЕЯ.

Современный (V) этап развития исследований и разработок в данной области характеризуется тем, что автоматической обработке подвергаются не искусственные (модельные) тексты, а реальные документы и, в общем случае, Web-контент; происходит обработка не единичных текстов, а мультиязычных коллекций документов; обрабатываемые документы содержат опечатки, орфографические ошибки, аграмматичности и другие реальные препятствия на пути к их правильной интерпретации. Кроме того, целью обработки документа становится не просто получение внутреннего представления его смысла, а представление результатов в форматах, удобных для эффективного хранения знаний с учетом их постоянного пополнения и последующего использования.

К сожалению, компьютерная лингвистика, искусственный интеллект и, тем более, информационные технологии не имеют на сегодняшний день мощных и эффективных моделей обработки текстов на естественных языках, а задача полной и правильной автоматической обработки произвольных ЕЯ-текстов, даже моноязычных, для произвольной предметной области пока не имеет сколько-нибудь практически значимого решения. Поэтому в настоящее время основное внимание исследователей и разработчиков сосредоточено на системах для извлечения информации из текстов (Information Extraction) – системах, названных по аналогии с разработкой месторождений, системами «разработки» текстов (Text Mining), а также на системах семантической классификации и кластеризации (Semantic Clustering/Classification) [21]. А наибольшую важность, как показывает анализ литературы [22,23], имеют IE-системы, обеспечивающие работу с мультиязычными коллекциями документов, которые могут быть получены из Интернет, новостных лент, блогов, корпоративных и персональных баз данных и других источников.

2.2.2. Карта леса

Анализ литературы и мониторинг информационных ресурсов Интернет показывает, что, как и следовало ожидать, наибольшая активность в данной области наблюдается в США и Канаде, Германии и Великобритании, за которыми следуют Италия, Франция и Япония. Имеются интересные коллективы и в других странах. Следует также отметить, что коллективы разных стран и даже одной страны существенно различаются по количеству сотрудников, квалификации и результатам. Так, в США имеются огромные исследовательские центры и корпорации, работающие в области обработки ЕЯ и создания соответствующих программных средств и систем (например, исследовательский центр ИБМ – Thomas Watson Research Center [24], лаборатория теории и технологий обработки корпорации Xerox – Palo Alto Research Center [25] или фирма Teragram [26]) и, наряду с этим, – небольшие исследовательские коллективы и фирмы, которые, тем не менее, разрабатывают интересные решения и полезные системы (например, группа обработки ЕЯ Стэнфордского университета [27] или американо-израильская компания ClearForest [28]).

Ситуация в Европе несколько отличается от ситуации в США и Канаде. Здесь, как правило, исследования и разработки ведутся в университетах, а результаты их «транслируются» в бизнес путем создания соответствующих start-up компаний. В Германии наиболее известным примером такого подхода является немецкий исследовательский центр ИИ (German Research Center for Artificial Intelligence) DFKI с центром речевых и языковых технологий [29] и образованная под его научным покровительством фирма ontoprise GmbH [30], в Великобритании – исследовательская группа обработки ЕЯ Шеффилдского университета (Natural Language Processing Research Group within the Department of Computer Science at the University of Sheffield) [31] и др.

Ситуация в России обсуждается во второй части настоящей работы, Поэтому здесь мы ограничимся по этому поводу лишь замечанием о том, что она отличается и от американской, и европейской.

2.3. Зарубежные разработки в области извлечения информации из ЕЯ-текстов

2.3.1. Вводные замечания

Поиск зарубежных исследовательских коллективов и компаний, позиционирующихся в области обработки ЕЯ, в Интернет показал, что даже «узкие» поисковые запросы с ключевыми словами и выражениями типа “information extraction”, “text mining”, “natural language processing systems” дают сотни тысяч ссылок, подавляющее большинство из которых мало информативно. Вот почему было принято решение перейти к обработке наиболее известных тематических каталогов и ресурсов, интегрирующих информацию по данной тематике. В результате обзора таких ресурсов были выбраны для предварительного анализа более 150 организаций (фирм, компаний, корпораций, университетских лабораторий и т.д.). Затем из этого перечня были намеренно исключены работы таких известных коллективов и организаций, как IBM Thomas J. Watson Research Center, Xerox Palo Alto Research Center, работы гренобльского отделения фирмы Xerox, а также работы других известных коллективов, известных достаточно широкому кругу специалистов в данной области. Вместо этого, на наш взгляд, интереснее провести анализ результатов, полу-

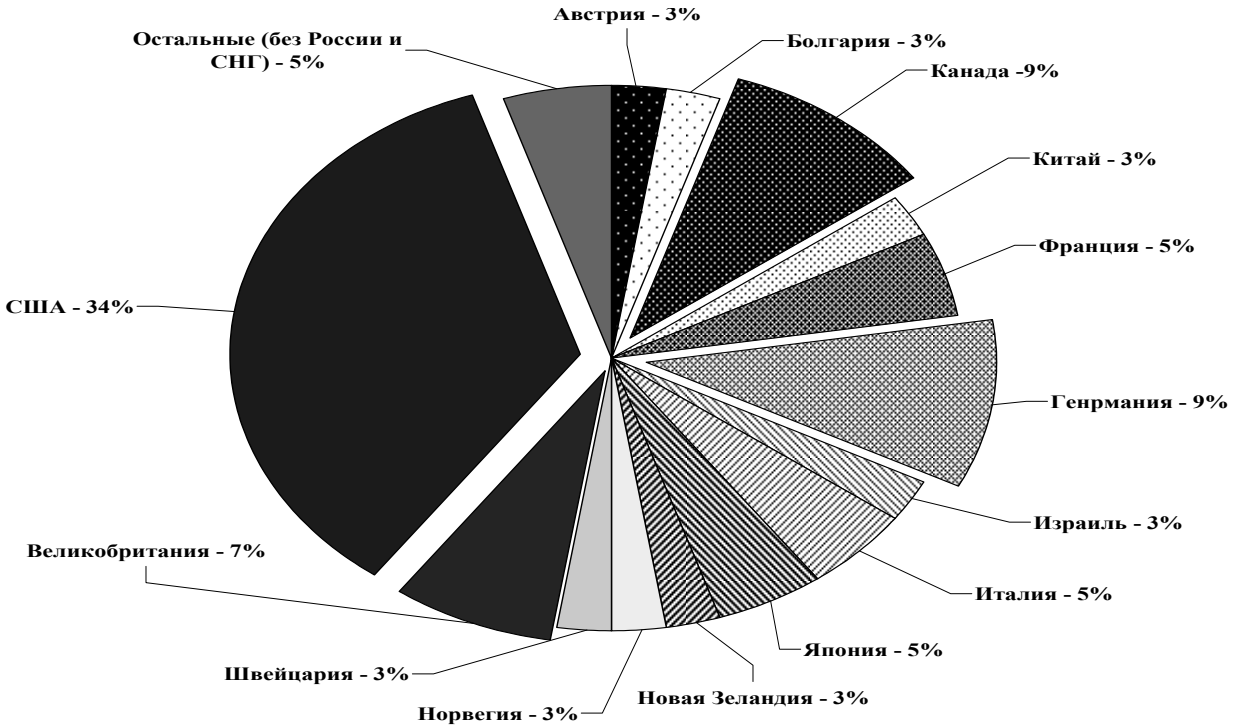

Рис. 2. Диаграмма территориального присутствия зарубежных исследовательских организаций и компаний, позиционирующихся в области NLP

ченных менее известными организациями, как исследовательскими, так и коммерческими. После предварительного анализа присутствия их в Интернете в списке было оставлено 23 организации. Основным критерием отбора на данном этапе было реальное функционирование соответствующих сайтов в Интернет с информацией, достаточной для проведения сравнительного анализа соответствующих разработок. Интегральные характеристики территориального присутствия зарубежных исследовательских организаций и компаний, позиционирующихся в данной области, с учетом замечаний сделанных выше, представлены на Рис. 2.

Следующим этапом стал более глубокий анализ отобранных сайтов, что привело к сокращению списка до следующих 14 организаций: BASIS Tech, ClearForest Corporation, CognIT, Compris Intelligence GmbH, Convera (formerly Excalibur), Delphes, Megaputer Intelligence Inc., Insightful Corporation & InFact, Inxight Software, Inc., MITRE, Ontotext, SRA International, Inc., TEMIS и Teragram Corporation [32,33,34,35,36,37,38,39,40,41,42,43,44,45].

2.3.2. Решения, продукты и системы

Для сравнения решений, продуктов и систем разных коллективов целесообразно, на наш взгляд, использовать такие основные критерии, как область охвата (спектр предполагаемых применений, обрабатываемых языков и функционалов, а также спектр пользователей) и технический уровень (поддержка стандартов, наукоемкость решений и их масштабируемость, а также тип решений – исследовательский прототип, экспериментальная система, тиражируемый продукт). Учитывая то, что эти критерии, в основном, качественные, ниже сначала описываются основные решения, продукты и системы, а затем проводится их сравнение в соответствии с указанными критериями.

Компания BASIS Tech

Из компаний, представленных в странах и регионах, отличных от США, Канады и Европы, после предварительного анализа, осталась для дальнейшего обсуждения единственная компания - BASIS Tech, которая имеет «двойное гражданство» (США, Япония).

Наиболее интересными из ее продуктов для нас являются платформа Rosette® Linguistics

Platform и инструментарий Arabic Desktop Tools.

Инструментарий Rosette® Linguistics Platform включает 2 основные компоненты: Rosette Base Linguistics и Rosette Entity Extractor. Rosette Base Linguistics поддерживает серьезный морфологический анализ, сегментацию и тэгирование текстов на китайском, японском, арабском, датском, английском, французском, немецком, итальянском и испанском языках.

Rosette Entity Extractor выделяет именованные сущности в мультиязыковых текстах на основе развитых технологий типа information extraction. Далее, на этой базе, осуществляется concept extraction (выделение концептов). Основные функционалы системы – Language Identifier, Поддержка Unicode, Base Linguistics, Name Matcher, Name Translator, Postal Address Analyzer. Основные сущности – Имена, Геообъекты, Организации, Даты. Из интересных свойств данной компоненты стоит отметить работу с номинативами, а также возможность обучения системы пользователями. Система работает с такими языками, как английский, французский, немецкий, итальянский, испанский, датский, португальский, русский, чешский, греческий, польский, венгерский, китайский, японский, корейский и арабский.

Решения компании Delphes

С момента образования компания разрабатывает высокопроизводительные масштабируемые системы типа Intelligent Knowledge Management для применения в области Enterprise Market. Все продукты группируются «вокруг» решений и функционалов, зафиксированных в Enterprise версии, где в процессе семантического поиска осуществляются выделение имен, субъектов, глаголов и т.п. сущностей для извлечения смысла запроса и соответствующих концептов; идентификация морфологии концептов; конвертирование чисел и распознавание таких концептов как даты, представленных в числовых и буквенно-цифровых форматах; распознавание таких концептов, как собственные имена, сложносочиненные слова (компаунды), акронимы, символы и аббревиатуры; идентификация синтаксической информации, необходимой для различения омонимов типа "leaves" (сущ.) и "leaves" (гл.); проверка правописания (Spell-check) запросов и автоматическая обработка вариантов; фиксация грамматических и орфографических ошибок; обеспечение подсказок в случаях неверной орфографии; определение языка документа и параграфа; различение «билингва» документов или «билингва» выражений в пределах одного документа. Enterprise версия поддерживает работу с английским, французским, испанским и немецким языками. Standard версия имеет «усеченный» функционал, исключена составляющая извлечения смысла. В дополнение к вышесказанному возможна кастомизация решений и интеграция их для таких языков как португальский, итальянский и японский, а также поставка специализированных словарей для корпоративного или личного использования.

Отдельным продуктом компании является Summarizer, выполняющий автоматическую генерацию релевантных рефератов по выбранной тематике на бизнес-портале заказчика. При этом соответствующий API обеспечивает возможности создания PDF, HTML или RDF версий рефератов, которые могут быть отправлены по электронной почте коллегам и бизнес-партнерам.

Компания Megaputer Intelligence Inc

Компания позиционируется в области создания и дистрибуции программных средств класса data mining, text mining и intelligent e-commerce personalization. В области text mining, которая находится в сфере интересов нашего анализа, Megaputer Intelligence предлагает решения на основе semantic network analysis с использованием методов класса Artificial Neural Networks для извлечения смысла из текста. Результаты такого анализа могут быть использованы для создания рефератов, concept-based searches, и даже для semantic text base navigation. Основные продукты Megaputer Intelligence в этой области – следующие: TextAnalyst 2.1, TextAnalyst for Microsoft Internet Explorer, TextAnalyst COM Objects и PolyAnalyst Qualitative Analysis Tools. В классе систем и решений Web mining компанией предлагаются продукты WebAnalyst и X-SellAnalyst.

В семействе систем TextAnalyst реализованы процедуры создания семантической сети обрабатываемого текста, которая создается полностью автоматически и не требует словарей и/или других априорных знаний. Для пользователя поддерживаются следующие функциональности:

Textbase Navigation (концепты, представленные в семантической сети текста, являются гипер-ссылками на те предложения, в которых они были найдены).

Topic Structure (система может идентифицировать наиболее важные концепты семантической сети и трансформировать сеть в дерево с упорядочиванием под-деревьев по важности).

Clustering (возможна оценка того, какие связи не важны, что позволяет «развалить» семантическую сеть текста на кластеры).

Summarization (семантическая сеть может использовать оценки отдельных предложений для получения текста контролируемой длины).

Natural language information retrieval (система определяет, содержатся ли слова запроса в семантической сети текста, после чего можно перейти к предложениям, в которых были найдены эти слова).

TextAnalyst доступен как отдельное приложение для MS Windows или как множество COM-компонент, которые могут быть легко интегрированы во внешние системы поддержки принятия решений. WebAnalyst является интеллектуальным сервером для e-Commerce и расширяет возможности существующего сервера за счет функционалов data и text mining.

Корпорация MITRE

В отличие от предыдущих компаний, корпорация MITRE является not-for-profit организацией, ориентированной на нужды федерального правительства США. В настоящее время MITRE управляет работой 3-х федеральных исследовательских центров: для министерства обороны и других правительственных организаций.

В интересующей нас области у корпорации имеется единственная разработка – инструментальная среда «The Alembic Workbench Environment for Natural Language Engineering». Основной функционал этой платформы концентрируется вокруг поддержки решения следующих задач:

- «Ручное» аннотирование текстов,
- «Ручная» композиция эвристик для IE в виде "phrase finding rule sequences",
- Автоматическое обобщение IE-правил с помощью машинного обучения,
- Оценка качества функционирования систем.

Платформа уже использовалась для аннотирования заголовков видео, тэгирования NE (people, organizations, locations) для португальских новостных текстов. В среде Alembic Workbench имеются мощные средства для быстрой разработки всех модулей типа pre-processing, part-of-speech tagging, phrase-tagging, а также синтаксического разбора для различных языков. Интересным решением является поддержка накопления, визуализации и анализа информации на уровне событий ("template entity" и "scenario template" в рамках MUC).

SRA International, Inc

Решения компании SRA в области text-mining ориентированы на широкий спектр организаций, работающих с большими объемами неструктурированных данных. В частности, основными прикладными областями использования продуктов фирмы являются Homeland Security, Enterprise Message Management, Enterprise Meta-Tagging и др.

Основная линейка продуктов компании SRA International, Inc. – NetOwl, первая версия которой была реализована в 1996 году. В настоящее время эта линейка включает NetOwl Extractor, v. 6, NetOwl DocMatcher, v. 2, NetOwl TextMiner, NetOwl InstaLink и NetOwl Summarizer.

NetOwl Extractor ориентирован на выделение именованных объектов из неструктурированных текстов на многих языках на основе использования методов компьютерной лингвистики. Продукт доступен в 2-х конфигурациях – NameTag и Link and Event, каждая из которых включает предопределенную онтологию. В конфигурации NameTag обрабатываются 7 основных и более 70 подтипов важных сущностей, в частности, people, organizations, places, artifacts, phone, social security numbers, dollar amounts, dates, addresses. В конфигурации Link and Event выделяются более 150 типов связей и событий на основе развитой онтологии. Примером выделяемых связей является связь типа affiliation (Who is affiliated with organization X?), примером события – transaction (Which country purchased satellites from company Y?). Основные преимущества данного решения в том, что здесь обеспечивается выделение не только объектов, но и связей и событий, в которые «вовлечены» соответствующие объекты; производится распознавание и классификация концептов с использованием лингвистического контекста ("Bush" person vs. "bush" plant; "Jor-

dan" place vs. person (e.g., "Michael Jordan); "fire" a weapon vs. "fire" a person); возможна кастомизация для поддержки извлечения других типов объектов, связей и событий с использованием конфигурации Creator Edition (CE); имеется поддержка возможностей «first entity translation», а также средства разрешения «алиасов» выделенных объектов с идентификацией их как ссылок на те же объекты реального мира ("FAA" => "Federal Aviation Administration", "the company's Chairman of the Board" => "John Smith").

Функционал NetOwl DocMatcher связан с определением «похожести» пар документов для идентификации дубликатов, полу дубликатов и концептуально схожих документов с использованием лингвистических атрибутов и алгоритмов машинного обучения. Данный продукт используется в задачах типа question-answer matching tasks, в приложениях типа intelligence analysis applications, которые осуществляют «мэппинг» новых документов на известную информацию, что необходимо для Customer Relationship Management (CRM), анализа патентов, поиска резюме.

NetOwl TextMiner предназначен для поиска, организации и анализа больших объемов неструктурированной информации. В нем интегрированы средства search engines, databases, visualization tools, вместе с развитыми инструментами text mining для извлечения информации, реферирования и кластеризации документов.

NetOwl InstaLink использует новые технологии уровня link analysis, visualization, и plan recognition для обеспечения доступа к смыслу связанной информации из различных источников. Основные свойства данного продукта – следующие: визуализация сложных сетей; автоматическое добавление новой информации из источников неструктурированных текстов с помощью техники drag-and-drop; возможность совместной работы аналитиков и поддержка многоязыкового интерфейса (включая арабский); распознавание вариантов написания имен, включая иностранные, а также хорошая масштабируемость по данным, поддержка популярных баз данных для накопления и поиска информации и 100% pure Java реализация, которая позволяет работать в режиме настольного приложения или запускаться с уровня Web браузера.

NetOwl Summarizer – генератор аннотаций и рефератов по длинным и сложным документам, использующий комбинацию лингвистических, статистических и обучающих техник, что делает его «аккуратным» и легко адаптируемым к новым типам документов. Основные свойства продукта – следующие: Theme-Based Summarization (определение основной темы каждого документа и генерация рефератов, правильно ее отражающих); Query-Based Summarization (возможность обучения для «фокусировки» рефератов на специальных требованиях пользователей); Adjustable Summary Length (возможность задания длины реферата); Adaptable for Different Text Types (легкая и быстрая адаптация к новым областям, контенту и источникам); простой и удобный API.

Оценивая продукты данной компании в целом можно констатировать, что она является серьезным разработчикам инновационных решений.

Teragram Corporation.

Teragram Corporation является OEM провайдером современных технологий типа Text, Linguistics и Information Extraction для всех основных европейских и азиатских языков. Основными продуктами корпорации являются TERAGRAM ENTITY EXTRACTION, TERAGRAM SUMMARIZER, а также EUROPEAN AND ARABIC LINGUISTIC SUITE.

Teragram Entities and Events Extractor автоматически выделяет сущности, концепты и события (например, people's and company's names, publicly traded businesses, titles and positions, and geographical locations). Концепты могут быть легко кастомизированы для специальных областей, заказчиков и использования. Продукт доступен как OEM-продукт для «обогащения» таких приложений как CRM, Knowledge Base, and Search, или как отдельные решения уровня предприятия. В дополнение к идентификации концептов в тексте Teragram Entities and Events Extractor может выдавать информацию, ассоциированную с ними (например, можно не только выделить publicly traded companies, упомянутые в документе, но и дать их ticker symbol и stock market, где они перечислены. The Teragram Entities and Events Extractor использует специальную технологию Teragram's Automatic

XML Marker technology для автоматического тэгирования выделенных концептов.

Teragram Summarizer позволяет создавать точные рефераты и даже рассылать их на мобильные телефоны. На основе технологий Teragram Linguistic and parsing создаются короткие рефераты с качеством уровня редактора-человека. Используемая технология реализована в разных архитектурах, масштабируема и доступна для европейских и азиатских языков.

Teragram EUROPEAN AND ARABIC LINGUISTIC SUITE. предоставляет технологии для Linguistic, information extraction, knowledge management и text processing для всех основных европейских языков, включая языки Северной и Восточной Европы.

ClearForest Corporation

В миссии компании явным образом декларируется, что ClearForest является поставщиком решений класса «text-driven business intelligence», что обеспечивает аналитический мост между двумя, прежде не связанными, информационными мирами – неструктурированными текстами и корпоративными данными.

Решения компании ClearForest ориентированы на следующие области применения: Quality Early Warning (обработка гарантийных обязательств с автоматическим извлечением из текстов гарантий таких элементов, как problem parts, failure conditions, technician service notes и/или comments), People and Corporate Profiles (извлечение информации из текстов типа business media и financial news таких элементов, как management changes, legal activities, mergers & acquisitions and products news), Federal Intelligence (сервисы для правительственных агенств, обеспечивающие выделение из текстов типа field reports, immigration records, Web content и emails такой информации, как associations between people and organizations, locations, weapon acquisitions), Patent Analysis (обработка патентных баз с целью выделения таких элементов, как key players, core patents, time-based chart of appearance).

Основным продуктом компании в интересующей нас области является платформа ClearForest Text Analytics Platform, ориентированная на интеллектуальный mark-up ключевых сущностей (person, organization, location), а также на фиксацию фактов или событий (например, идентификация сущностей, характерных для определенной индустрии и связи этих сущностей с другими) в свободных текстах (например, в новостях, Интернет-обзорах и HTML-документах). Платформа имеет два основных компонента: ClearForest Tagging Engine и ClearForest Extraction Modules. Вместо описания функционалов этих компонент в подавляющем большинстве доступных материалов делаются лишь рекламные заявления о их мощности и указываются области применения. Вместе с тем, анализ научных статей, подготовленных специалистами ClearForest для различных конференций, позволяет сделать вывод о том, что данными модулями обеспечивается следующий перечень функционалов: морфологический и лексический анализ (на уровне POS-tagging и снятия лексической омонимии); синтаксический анализ (поверхностный и глубинный) и семантический анализ (включая обработку анафоры и интеграцию результатов).

При этом сущности (например, person или organization) обрабатываются с «аккуратностью» (непонятно, что имеется ввиду, но можно надеяться, что это F-мера) 90-98%; атрибуты (свойства сущностей, такие как названия, алиасы и т.п.) с «аккуратностью» 80%; факты (отношения между сущностями, такие как, Position of a Person in a Company) с «аккуратностью» 70-90% и события (активности, в которые вовлечены сущности, например, terrorist act, airline crash, management change, new product introduction) с «аккуратностью» 60-80%. Точного перечня обрабатываемых объектов нет. С другой стороны, ClearForest, совместно с компанией Celera Genomics (NYSE:CRA), был победителем 2002 Knowledge Discovery and Data Mining (KDD) Cup, где анализировались научные статьи по теме Drosophilia Fruit Fly.

В 2008 году компания (а вернее ее часть) была куплена информационным агентством «Рейтер», которое предполагает использовать технологии обработки ЕЯ-текстов ClearForest для семантизации новостного контента. На портале компании сейчас функционирует бэта версия сервиса Gnosis, обеспечивающая обработку небольших текстов в режиме «on fly». Кроме того, компания открыла API для обращения к сервисам обработки ЕЯ-текстов на английском языке. Первые результаты тестирования этого сервиса, представленные на Рис. 4, показывают, что качество обработки текстов

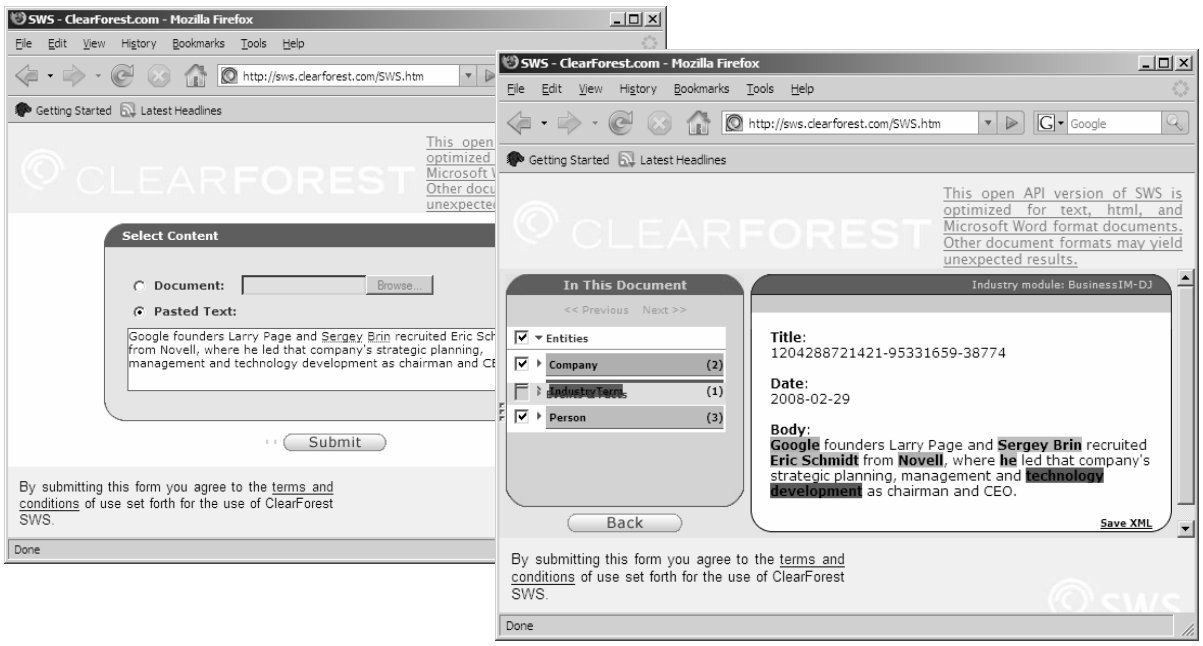

Рис.4. Результаты тестирования сервиса Gnosis

далеко не всегда соответствует заявлениям разработчиков, хотя сервис активно развивается и пополняется.

Компания CognIT

Продукты и решения компании CognIT концентрируются вокруг следующих основных областей применения: инновационные машины поиска, Content Management Support (CMS) и Knowledge Management and Best Practice (KM). Если говорить более конкретно, продукты компании, разработанные на основе технологий, созданных исследовательскими лабораториями Норвегии, зарегистрированы под маркой CORPORUM® и могут использоваться для интеллектуального поиска и индексирования, структуризации контента на порталах, аннотирования документов по контенту, реферирования и сжатия информации, а также для извлечения имен и отношений из текстов.

Линейка продуктов CORPORUM® Intranet Search & Navigation включает CORPORUM® SLATEWEB и CORPORUM® engine. Инсталляция CORPORUM® SLATEWEB поступает к пользователю с несколькими готовыми для использования приложениями, такими как web crawler, file indexer и др. В нее входят также анализаторы для нескольких форматов документов (в частности, PDF и Microsoft Word). И, наконец, SLATE WEB включает последнюю версию CORPORUM® engine, поддерживающего английский, норвежский, немецкий и шведский языки.

Основные функционалы SLATEWEB связаны с поиском (точные образцы и множественные термы, полная поддержка булевских выражений, неограниченная вложенность скобок), сортировкой (по релевантности и дате), интервальным поиском (фильтрация результатов по дате, релевантности и предметной области), лингвистическим анализом (автоматическое определение языка документа, NLP, реферирование и извлечение концептов, сущностей, связанных концептов), таксономией и классификацией (категоризация на основе правил), автоматической генерацией таксономии, а также поддержкой краулинга по многим источникам (Internet, Intranet, IMAP, Exchange, Oracle, MSQL) и с некоторыми другими функционалами.

В основе CORPORUM® engine лежит технология CognIT Mimir, с помощью которой тексты интерпретируются на основе онтологий, отражающих модель интересов пользователя, а сама модель интересов рассматривается как база знаний для определения контекстного и тематического соответствия обрабатываемых до-

кументов этой модели. При этом модель интересов пользователя управляет процессом извлечения информации, а результаты сохраняются в базе данных для дальнейшего использования. Для представления и обработки знаний используется специальная среда The CORPORUM® portfolio, включающая, по последним данным, CORPORUM® Knowledge Factory, CORPORUM® Knowledge Server, CORPORUM® Summarizer и CORPORUM® Best Practice. Более развернутая информация о детальных характеристиках отдельных составляющих CORPORUM® portfolio практически недоступна.

По представленной компанией информации ее продукты производят хорошее впечатление. Однако существует опасность, что их мощность в интересующей нас области меньше, чем это кажется из перечисления их свойств.

Compris Intelligence GmbH

В миссии компании явным образом декларируется, что Compris Intelligence позиционируется в области «text understanding technologies». При этом собственно этой миссии отвечают только несколько продуктов, а остальные (часто интересные) находятся, скорее, в сфере нетривиальной обработки данных.

Учитывая вышесказанное, ниже рассматриваются два продукта фирмы Compris Intelligence: Information extraction (FactMiner) и Language components.

Основные свойства продукта FactMiner – следующие: обрабатываются документы типа (Text, RTF, HTML, SGML, XML, PDF, PostScript); из текстов выделяются адреса (contact persons, candidates for jobs, responsible persons for Web servers), информация о продукте (product name, product description, listing of the properties/features, price, availability); автоматический сбор информации о стилевых особенностях текста и лингвистических преференциях для автоматического расширения лексики; автоматическое обучение на параллельных текстах на разных языках; распознавание, классификация и интеллектуальное распределение релевантных сообщений (news, e-mail, mailing list contributions, internet news). Извлечение информации из текстов ведется на гибридных моделях (комбинациях правил с вероятностями их применения), что обеспечивает корректную обработку неоднозначностей при выделении сущностей (например, немецкие имена "Rudolf", "Dieter", "Thomas" могут быть как первыми именами, так и фамилиями, а использование вероятностей помогает выбрать корректный вариант). В системе реализовано мощное распознавание адресов (например, адреса могут быть распределены по нескольким строкам/колонкам таблицы; наименования улиц и городов трудно выделяемые в обычных системах здесь обрабатываются правильно; возможность интеграции адресной информации (с одного сайта общий адрес, а с другого – кому именно адресовать ту или иную информацию); включение в адрес информации из графических логотипов, и т.п.).

Продукты линейки Language components доступны в виде соответствующих инструментальных пакетов (SDK) или в виде наборов данных (data sets). В настоящее время это следующие ресурсы: морфологический процессор, не зависящий от языка (Finite State Morphology), реализованный на языке C++; морфологические данные для немецкого и английского языков (около 50 000 базовых форм для каждого); словари синонимов для немецкого и английского языков (с отношениями типа is-a, part of, element of, antonym of), при этом большинство синонимов снабжены контекстной информацией (например, значение слова "bank" может быть отделено от других значений этого же слова логическими связками вида OR/AND/NOT; средства сокращения неоднозначностей за счет трансформации фрагмента текста в фразовую структуру (phrase structure) для немецкого и английского языков; генерация стандартных формулировок из множества возможных путем замены синонимов и изменения порядка слов.

В целом продукты этой фирмы не производят сильного впечатления, хотя несколько интересных свойств в них имеется.

Convera (formerly Excalibur)

В интересующей нас области у компании Convera обозначена линейка продуктов, в состав которой входят лингвистические процессоры (Language Processors) и картриджи знаний (Knowledge Cartridges). Информация об этих продуктах представлена крайне скудно и, по преимуществу, не в техническом плане.

Лингвистические процессоры компании Convera обладают развитыми средствами обра-

ботки ЕЯ, включая language identification, tokenization, morphology analysis, idiom processing и part-of-speech tagging. Кроме того, Convera использует специальные средства подготовки данных, которые обеспечивают нормализацию, удаление stop-word и идентификацию фраз.

Картриджи знаний RetrievalWare обеспечивают масштабируемые и гибкие мультиязыковые технологии для управления знаниями на основе использования тезаурусов и контролируемых словарей (controlled vocabularies), таксономий и списков сущностей. Convera поставляет более 70 различных типов картриджей, объединенных в следующие 5 категорий:

Domain-specific cartridges (тысячи специфических для разных предметных областей взаимосвязанных терминов).

General semantic cartridges (словари, организованные как семантические сети. Такие словари опираются на морфологию, а также имеют средства обработки идиом. Некоторые из картриджей – мультиязыковые).

Taxonomy cartridges (60 таксономий, каждая из которых содержит тысячи категорий на 10 уровнях иерархии. Таксономии RetrievalWare построены на основе индустриальных стандартов в области тезаурасов и опираются на ресурсы Medical Subject Headings (MeSH), Defense Technical Information Center (DTIC), Proquest and WAND).

Entity cartridges (в частности, списки известных организаций (подразделений, партнеров), продуктов, людей (включая ссылочную информацию (фамилии, телефоны, адреса), Web-ресурсы (URLs, e-mail и IP addresses). Кроме того, Entity Cartridges могут содержать правила и образцы, которые используются для поиска соответствующих понятий в текстах.

Cross-lingual cartridges (кросс-языковые картриджи).

Insightful Corporation & InFact

В интересующей нас области у компании Insightful Corporation обозначен единственный продукт InFact. Техническая информация об этом продукте практически недоступна, хотя из разрозненных статей можно констатировать, что это решение уровня text mining для информационного поиска. Компания утверждает, что система содержит 20 подсистем, большинство из которых изначально были разработаны для военного ведомства США. Утверждается также, что InFact способен понимать текст на уровне, близком к человеческому, и модифицировать свое поведение в зависимости от полученных результатов. InFact в настоящее время доступен для компьютеров, работающих под управлением Sun Solaris systems. Стоимость лицензии InFact для одной системы управления знаниями начинается с $250,000.

Inxight Software, Inc

У компании Inxight имеется серьезная линейка продуктов по обработке ЕЯ. Продукты доступны как в виде части серверных решений, так и в виде инструментальных сред (SDK). Открытый и гибкий API и XML-выход позволяет интегрировать Inxight-технологию в другие приложения. Линейку интересующих нас продуктов можно фиксировать следующим образом: продукты анализа текстов (Text Analysis Products), продукты визуализации (Visualization Products) и инструментарий (SDK).

Компания Inxight обладает уникальным набором инструментов для обработки текстов на всех основных языках. В этом разделе компания позиционирует 2 продукта: Inxight SmartDiscovery Extraction Server™ и SmartDiscovery Analysis Server™ . Продукты используются для категоризации, поиска и приложений класса business intelligence.

Продукты визуализации Inxight (StarTree®, TableLens® и TimeWall™) позволяют пользователям выявлять скрытые отношения и тренды на больших объемах информации. Продукты доступны в составе Inxight VizServer® или как отдельные Software Development Kits (SDKs).

Компоненты ядра Inxight также доступны в составе Software Development Kits (SDKs). Они легко интегрируются с другими прикладными системами.

Inxight SmartDiscovery Extraction Server™ - мощная линейно масштабируемая система, которая может обрабатывать терабайты данных. При этом используются функциональные возможности системы типа IE ThingFinder®, которая совместно с модулями Categorizer и Summarizer позволяет структурировать информацию из неструктурированных текстов. Система позволяет пользователям выделять более 35 типов сущностей, обрабатывать сущности на базе технологии образцов и/или через списки сущностей, определять события и отношения,

импортировать или разрабатывать собственные таксономии на базе обучения по примерам или на основе точных лингвистических правил, получать интеллектуальные рефераты.

Совокупность компонент, входящих в Inxight LinguistX® Platform, обеспечивает автоматическую идентификацию языка и кодировку слов документа, а также анализ документов, включая идентификацию параграфов, нормализацию регистров, сегментацию слов, стемминг, декомпозицию сложных слов, POS-тэгинг, а также выделение именных групп. Сейчас доступны соответствующие модули для следующих языков: арабский, китайский (упрощенная и традиционная нотация), датский, английский, фарси (персидский), французский, немецкий, итальянский, японский, корейский, португальский, русский, испанский, шведский, финский и др.

Кроме рассмотренных в линейке продуктов Inxight имеется Summarizator, правда, как следует из доступных документов, работающий на традиционных идеях.

Ontotext

Говорить о продуктах компании Ontotext в интересующей нас области особого смысла нет. Вместе с тем, специалисты компании сделали серьезный вклад в разработку и реализацию компонент платформы GATE, которые могут быть интересны в нашем случае. Учитывая вышесказанное, дадим краткую характеристику этих компонент.

Japec (JAPE-to-Java compiler), ресурс GATE 3.1+. Данный компилятор слабее, чем компилятор Jape+, реализованный российской компанией Авикомп Сервисез по мощности входного языка, но, по-видимому, быстрее. Можно предположить, что в настоящее время идет интенсивное расширение входного языка.

GATE's Oracle support – поддержка хранилища данных ORACLE, ориентированная на работу с большими корпусами текстов.

Optimizations – оптимизатор модулей GATE, что дало повышение производительности больше, чем в 2 раза.

WordNet API – интеграция лексической базы WordNet с GATE через Java API.

Protege-2000 integration – интеграция в GATE известного инструментария онтологического инжиниринга.

Ontology access, editing, and markup – поддержка редактирования и использования онтологий на уровне хранилища DAML+OIL с доступом через API, что позволяет использовать иерархические аннотации.

Кроме того, специалистами компании разработаны компоненты, которые не поставляются с открытым кодом. В частности, это Hash Gazetteer (работает с хэш-таблицами вместо FSM и обеспечивает 4-х кратное уменьшение памяти при 3-х кратном ускорении. Аналогичная подсистема разработана специалистами российской компании Авикомп Сервисез несколько лет назад. Hidden Markov Model Learner – стохастический модуль для фильтрации аннотаций и снятия неоднозначностей на основе «мер доверия».

В области обработки естественного языка данная компания собственных достижений практически не имеет, но, вместе с тем, у нее имеются серьезные результаты в разработке RDF-хранилищ, а также создания прикладных систем для Semantic Web.

TEMIS

Продукты компании TEMIS ориентированы на работу в таких приложениях, как Life Sciences, Energy/Utilities, Publishing, Homeland Security, Automotive.

С точки зрения целей настоящего анализа интересны такие продукты компании, как Insight Discoverer™ Extractor, Insight Discoverer™ Clusterer и XeLDA®. На сайте компании все они представлены фрагментарно, однако анализ литературы позволяет дать характеристики этих продуктов, представленные ниже.

Insight Discoverer™ Extractor предназначен для извлечения информации из текстов типа e-mail для CRM, специальные навыки из резюме и т.п. Подход к разработке использует итеративные методы: сначала проводится морфосинтаксический анализ, затем распознавание сущностей, а затем выделение образцов ситуаций на основе лингвистических и семантических помет. Используемые правила достаточно очевидны и безусловно не покрывают всего разнообразия возможных конструкций. Основное достижение – использование больших корпусов текстов по заданной тематике в предположении, что в них всегда найдутся фрагменты, в которых «сработают» простые правила.

В отличие от предыдущих продуктов лингвистический инструментарий XeLDA (правда,

разработанный в Xerox) оставляет самое серьезное впечатление. Основные функционалы его – следующие: Language identification (распознавание языка, на котором написан текст), Tokenization (деление текста на слова), Morphological analysis (морфологический анализ), Part of speech disambiguation (определение грамматических категорий по контексту), Noun phrase extraction: (идентификация именных групп), Dictionary lookup (использование контекста для разрешения словарной омонимии), Idiom recognition (распознавание идиом), Relational morphology (группирование слов в соответсвии с их дериватами). Платформа поддерживает большинство западно-европейских и некоторые из восточно-европейских языков. Языки Среднего Востока в активной разработке. Конкретно, в настоящее время поддерживаются чешский, датский, английский, французский, немецкий, греческий, венгерский, итальянский, польский, португальский, русский и испанский. Используемые словари первоначально кодируются в XML или SGML формате, а затем компилируются в формат XeLDA. Вместе с тем, морфология реализована для английского, французского, немецкого, итальянского и испанского языков. Серьезным модулем является модуль обработки идиом. При этом возможна замена отдельных слов идиоматического выражения на их эквиваленты на другом языке.

В целом данная компонента оставляет серьезное впечатление по компонентам, но не дает оснований считать, что в ее рамках есть хорошая поддержка этапов дальше морфологии и поверхностного синтаксиса.

2.3.3. Оценка решений, продуктов и систем зарубежных разработчиков

Приведенная выше информация о решениях, продуктах и системах разных коллективов зарубежных разработчиков позволяет высказать некоторые, на наш взгляд, небезынтересные соображения и оценки.

Во-первых, большинство коллективов ориентировано на широкий спектр обрабатываемых и/или поддерживаемых языков (16 у BASIS Tech, 15 у Inxight, 12 у TEMIS, все основные европейские языки у Teragram). Вместе с тем, широкий охват далеко не всегда означает глубокую наукоемкую проработку (так, например, компания Convera имеет громадные словари, неплохие морфологии для серьезного спектра ЕЯ, но использует достаточно слабые лингвистические модели следующих за морфологией уровней). С другой стороны, компании и коллективы, ориентированные на небольшой спектр обрабатываемых языков, как правило, имеют серьезные результаты в области их синтаксического анализа и, правда реже, на более глубоких лингвистических уровнях. Часть организаций явно позиционируется в области обработки мультиязыковых документов.

Во-вторых, практически все представленные в настоящем обзоре организации имеют достаточно широкий спектр средств предварительной обработки текстов. Как правило, это определение языка документа, коррекция опечаток, работа с разными форматами входных документов и т.п.

В-третьих, некоторые из представленных в настоящем обзоре коллективов и организаций имеют серьезные результаты не только в области обработки ЕЯ, но и создают мощные инструментальные средства поддержки лингвистических разработок. Здесь, на наш взгляд, явно выделяются такие компании, как BASIS Tech с платформой Rosette® Linguistics Platform и инструментарием Arabic Desktop Tools, Inxight и его платформа LinguistX® Platform, Ontotext, традиционно ориентированный на GATE, MITRE с инструментальной средой «The Alembic Workbench Environment for Natural Language Engineering»TEMIS с его инструментарием XeLDA, а также Teragram с лингвистическим инструментарием EUROPEAN AND ARABIC LINGUISTIC SUITE.

Наконец, «номенклатура» функционалов, продуктов и сервисов, разрабатываемых коллективами и организациями, представленными в настоящем обзоре, концентрируется, в основном, вокруг выделения из текстов достаточно ограниченного набора именованных сущностей. И только некоторые из них, во-первых, имеют дело с серьезным спектром объектов (Inxight – более 35 типов сущностей, SRA International – 7 основных и более 70 подтипов важных сущностей), а во-вторых, идут на выделение из текстов семантических отношений между сущностями и/или событий и других артефактов, связанных с объектами (SRA International, Teragram). Из прикладных функционалов достаточно «частотными» являются реферирование (CognIT, Delphes, Megaputer

Intelligence, SRA International, Teragram), а также кластеризация и классификация (CognIT, Convera, Megaputer Intelligence, Ontotext, TEMIS, SRA International). Несколько коллективов и организаций, представленных в настоящем обзоре, имеют серьезные разработки в области семантической навигации и визуализации (ClearForest, Compris Intelligence, Inxight, Megaputer Intelligence, Ontotext, SRA International).

Результаты проведенного сравнительного анализа решений и разработок коллективов и организаций, представлены в таблице.

Заключение

В первой части настоящей статьи представлено активно развивающееся на стыке работ в области искусственного интеллекта и Интернет-технологий новое направление – Семантический Вэб. Дан аналитический обзор состояния исследований в данном направлении, в частности, краткая история вопроса, фундаментальные проблемы организации пространств знаний в сети Интернет и возникающие в связи с этим задачи извлечения знаний из текстов на естественных языках.

Более подробно в этой части статьи обсуждались методы и средства извлечения информации из текстов на примере анализа решений, продуктов и систем, разрабатываемых зарубежными коллективами и организациями, работающими в этой области.

В следующей части данной работы предполагается провести обзор решений, продуктов и систем, разрабатываемых в России и странах СНГ, а также остановиться более подробно на вопросах семантической навигации по пространствам знаний и прикладных интеллектуальных системах, ориентированных на Семантический Вэб.

Литература

1. http://www.idc.com/research/reshome.jsp
2. Alvin Tofler. Third Wave, Bantam Books, 2006, ISBN 0-553-24698-4
3. Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001 (http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21)
4. Mills Davis. Semantic Wave 2006. Executive Guide to Billion Dollar Markets. A Project10X Special Report. January 2006.
5. Cearley, D. W., Andrews, W., Gall, N.: Finding and Exploiting Value in Semantic Technologies on the Web. 9 May 2007, ID №: G00148725. Gartner, Inc. (2007)
6. V. R. Benjamins, J. Contreras, O. Corcho and A. Gomez-Perez, Six Challenges for the Semantic Web, http://www.cs.man.ac.uk/~ocorcho/documents/KRR2002WS_BenjaminsEtAl.pdf
7. Tim Berners-Lee, The Semantic Web and Research Challenges, http://www.w3.org/2003/Talks/01-sweb-tbl/slide1-0.html
8. Michael Kifer, Georg Lausen, James Wu. Logical Foundations of Object Oriented and Frame Based Languages. Journal of ACM 1995, vol. 42, p. 741-843
9. Jeff Heflin, James Hendler, and Sean Luke. Applying Ontology to the Web: A Case Study, In International Work-Conference on Artificial and Natural Neural Networks, IWANN'99. 1999.
10. http://www.daml.org/2001/03/daml+oil-index.html

Коллективы, организации	Спектр обр. языков	Средства предв. обработки	Инстр.	Спектр выд. объектов (отн.)	Рефер.	Кластер., классиф.	Сем. навиг. и визуал.
BASIS Tech	16	+	+	+(-)			
ClearForest	1	+		+(+)			+
CognIT	1+?	+		+(-)	+	+	
Compris Intelligence	2+?	+		+(-)			+
Convera	много	+		+(-)		+	
Delphes	4	+		+(-)	+		
Megaputer Intelligence	1	+		+(-)	+	+	+
Insightful Corp. & InFact	1+?	+		+(-)			
Inxight Software	15	+	+	35(+)			+
MITRE	1+?	+	+	+(-)			
Ontotext		+	+	+(+)	+	+	+
SRA International	много	+		7 осн. и 70 подтипов (+)	+	+	+
TEMIS	12	+	+	+(-)		+	
Teragram	Все евр.	+	+	+(-)	+		

11. http://en.wikipedia.org/wiki/Owl
12. Хорошевский В.Ф., Информационное пространство РАИИ в среде Internet, Труды V национальной конференции с международным участием "Искусственный Интеллект-96", Казань 5-11 октября 1996 г., Центрпрограммсистем, Тверь, 1996.
13. Khoroshevsky V.F., Knowledge vs Data Spaces: How an Applied Semiotics to Work on Web, In: Proceedings "3rd Workshop on Applied Semiotics", National Conference with International Participation (CAI'98), Pushchino, Russia, 1998.
14. Benjamins R., Decker S., Fensel D., Gomez-Perez A. "$(KA)^2$: Building Ontologies for the Internet": A Mid Term Report. International Journal of Human Computer Studies (IJHCS). 51(3). September 1999.
15. Decker, S.; Erdmann, M.; Fensel, D.; Studer, R. Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information. In R. Meersman et al. (eds.): Semantic Issues in Multimedia Systems. Proceedings of DS-8. Kluwer Academic Publisher, Boston, 1999
16. Alexander Maedche, Steffen Staab: "Learning Ontologies for the Semantic Web", Semantic Web WorkShop 2001, Hongkong , China, 2001
17. http://www.w3.org/RDF
18. http://www.w3.org/2004/
19. http://en.wikipedia.org/wiki/Microformats
20. В.Ф. Хорошевский, Обработка естественно-языковых текстов: от моделей понимания языка к технологиям извлечения знаний, Журнал «Новости ИИ», № 6, 2002.
21. Proceedings of the Twelfth Text Retrieval Conference (TREC 2003), http://trec.nist.gov/pubs/trec12/
22. T. Poibeau, A. Acoulon, C. Avaux, L. Beroff-Bénéat, A. Cadeau, M. Calberg, A. Delale, L. De Temmerman, A.-L. Guenet, D. Huis, M. Jamalpour, A. Krul, A. Marcus, F. Picoli and C. Plancq, The Multilingual Named Entity Recognition Framework, In the EACL 2003 Proceedings (European Conference on Computational Linguistics), Budapest, 15-17 April 2003.
23. Proc. 4th International Conference On Language Resources And Evaluation (LREC 2004), Lisbon, Portugal, 26-28 May 2004.
24. http://www.research.ibm.com/talent/members.html
25. http://www2.parc.com/isl/groups/nltt/
26. http://www.teragram.com/
27. http://nlp.stanford.edu/
28. http://www.clearforest.com/index.asp
29. http://www.lt-cc.org/index-e.html
30. http://www.ontoprise.de/content/index.html
31. http://nlp.shef.ac.uk/
32. http://www.basistech.com/
33. http://www.delphes.com
34. http://www.megaputer.com
35. http://www.mitre.org
36. http://www.sra.com
37. http://www.teragram.com
38. http://www.clearforest.com/
39. http://www.cognit.no
40. http://www.kompass.com
41. http://www.convera.com
42. http://www.insightful.com
43. http://www.inxight.com/
44. http://www.ontotext.com
45. http://www.temis.com

Хорошевский Владимир Федорович, зав. отделом 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 Вэб.