Система PullEnti – извлечение информации из текстов естественного языка и автоматизированное построение информационных систем*

***О.В. Золотарёв*[1], *М.М. Шарнин*[2], *С.В. Клименко*[3], *К.И. Кузнецов*[2]**
ol-zolot@yandex.ru|mc@keywen.com|stanislav.klimenko@gmail.com|k.smith@gmail.com
[1]Москва, АНО ВО Российский новый университет
[2]Москва, Федеральный исследовательский центр «Информатика и управление» РАН
[3]Протвино, Институт физико-технической информатики

В работе исследуются вопросы анализа текстов естественного языка на основе технологии PullEnti. Из текстов (документов) извлекаются интересующие пользователя объекты, их свойства и связи. Представляются факты участия объектов в действиях. Рассматриваются технологии обработки больших массивов информации, методы построения информационных систем на основе предложенной технологии, анализируются сложности при анализе текстов естественного языка. Используемый подход будет проработан на примере анализа текстов по компьютерной графике на предмет выявления неявных ссылок, и эти наработки будут использованы при анализе террористической активности в сети интернет.

Ключевые слова: *лингвистический процессор, обработка естественного языка, извлечение объектов, именованные сущности, связи объектов*

System PullEnti – Information Extraction from Natural Language Texts and Automatic Construction of Information Systems*

***O.V. Zolotariev*[1], *M.M. Charnine*[2], *S.V. Klimenko*[3], *K.I. Kuznetsov*[2]**
[1]Moscow, Russian new university
[2]Institute of Informatics Problems FRC CSC of RAS, Moscow
[3]Institute of Computing for Physics and Technology, Protvino

This paper investigates the analysis of natural language texts based on the technology PullEnti. Objects, their properties and relationships are retrieved from natural language texts. The article discusses the technology of processing of large amounts of information, methods of construction of information systems based on the proposed technology. It also analyzes the complexity in the analysis of natural language texts. The approach will be elaborated on the example of analysis of texts in different spheres. It also analyzes natural language texts on the Internet. The paper presents examples of the results of the semantic analysis of texts.

Keywords: *linguistic processor, natural language processing, data extraction, named entities, relationships, objects*

Введение

Способы представления информации, знаний многообразны. Огромное количество информации представлено в виде текстов естественного языка, что делает задачу извлечения и структурирования информации из текстов весьма важной. Это относится к различным предметным областям. Для оперирования информацией в компьютере необходимо выделить из текста объекты, их атрибуты, связи между объектами, процессы, в которых эти объекты задействованы, другую важную информацию, которая бы позволяла не только описать ситуацию, но и строить выводы, характерные для конкретной предметной области, прогнозировать развитие ситуации.

Существует большое количество задач, которые могут решаться на основе анализа текстовой информации. Это и задачи связанные с определением уровня террористической угрозы в сети интернет, и задачи поиска групп террористов, разного рода задачи кластеризации текстов для сужения пространства поиска, задачи определения неявных ссылок между документами, в тех случаях, когда они ссылаются на одни и те же идеи, и задача построения моделей бизнес-процессов, и многие другие.

Выделение неявных ссылок в документах требует проведения анализа фраз, встречающихся в данном документе и выявления идей, которые могут появляться в других документах. Это позволяет проводить анализ цитируемости статей, монографий не только на основе использования явных библиографических ссылок, но и посредством неявных ссылок, которые определяются, когда одна и та же идея встречается в разных документах. Проблема состоит в том, чтобы выявить источник появления идеи. Это может быть сделано следующим

Работа выполнена и опубликована при финансовой поддержке РФФИ, гранты 16-07-00756, 16-07-20854, 16-29-09527, 15-07-06586, 16-37-50057

образом: отыскивается документ, в котором данная идея официально была опубликована раньше, чем в других документах. В случае невозможности определения даты самой ранней публикации документа, в котором встретилась искомая идея, ее автор может быть выбран на основе максимального количества появлений данной идеи в опубликованных документах.

Построение коллекций документов на основе анализа информации в сети Интернет

Работа программы PullEnti может быть проверена на массиве текстов. Для подготовки таких тематических массивов была разработана технология построения коллекций по открытой информации из Интернета.

В частности, для целей данного исследования была разработана специальная технология, которая позволяет строить коллекции научных публикаций для любой предметной области из открытых источников в Интернете. Массивы документов могут быть построены в форматах XML-, TXT- и PDF.

Для построения массива документов была использована модифицированная Keywen-технология построения интернет-корпусов [1, 2, 3, 4, 5, 6, 7], первоначально разработанная для автоматизированного построения интернет-энциклопедии `Keywen.com`. Данная технология аналогична широко известным методам построения интернет-корпусов
ГИКРЯ (webcorpora.ru),
НКРЯ (ruscorpora.ru)
и ruTenTen (sketchengine.co.uk).
Для целей данного исследования Keywen-технология была настроена на поиск в Интернете PDF-файлов, содержащих название, авторов, списки литературы, а также термины из области Компьютерной Графики. Keywen-технология основана на собственном поисковом роботе, который ищет тексты, ассоциированные с заданными ключевыми терминами. Ниже приведен список ключевых терминов, использованный в данном исследовании при формировании коллекции документов по тематике «компьютерная графика»:

компьютерная графика, виртуальная реальность, научная визуализация, геометрическое моделирование, машинная графика, визуальный анализ, методы визуализации, визуализация поверхности, пространственная сцена, пространственное моделирование, многоугольная фигура, системы виртуальной реальности, авиационные тренажеры, визуальная аналитика, визуализация информации, осязаемые изображения, 3D-визуализация, виртуальная среда, графический процессор, виртуальная сцена, очки виртуальной реальности, шлем виртуальной реальности, анализ изображения, 3D-ландшафт, моделирование территории, векторная графика, растровая графика, фрактальная графика, трёхмерная графика, 3DMAX, визуализация данных, трёхмерная визуализация, объемная визуализация, компьютерная визуализация

Найденные при помощи Keywen-технологии PDF-файлы были переведены в текстовый формат при помощи программы FineReader и далее были преобразованы в формальный XML-формат при помощи специально разработанного лингвистического процессора, который выделял из текста название, авторов и библиографические ссылки. Разработанная в данном исследовании модифицированная Keywen-технология позволяет строить коллекции научных публикаций для любой предметной области из открытых источников в Интернете.

Также разработаны механизмы пополнения коллекции, автоматически добавлено более 900 публикаций. Описания публикаций включают метаданные, связи с другими объектами БД, текстовый слой, а также PDF-файл с полным текстом

PullEnti – программа для анализа текстов естественного языка

Для выделения идей и ссылок на них необходимо, во-первых, иметь доступ к библиотекам документов по различным сферам деятельности, во-вторых, требуется производить автоматическую обработку больших массивов текста на предмет выделения неявных ссылок, т.е. ссылок на идею.

Именно для этой цели в работе использован лингвистический процессор PullEnti. Выбор этого программного продукта обусловлен рядом факторов. С одной стороны, данный процессор разрабатывается в сотрудничестве с институтом Проблем информатики, в котором и работают авторы статьи, с другой стороны, этот процессор в рамках проводимых соревнований конференции Диалог-2016 занял несколько первых мест при анализе текстов в рамках решения различных задач. Разработчик данного процессора – Кузнецов Константин Игоревич.

PullEnti представляется из себя программный пакет, включающий алгоритмы морфологического и синтаксического анализа, который позволяет выделять сущности определенных типов из текстов естественного языка (персоны, организации, ...). Именованная сущность – это объект, содержащий набор значений атрибутов, отличающий его от других объектов этого же типа.

В системе PullEnti используются динамически подключаемые плагины, что позволяет без перекомпилирования подключать различный функционал. Именно таким образом подключается блок семантического анализа.

В процессе анализа выделяются так называемые токены, которые представляют собой типизированные фразы, такие как текстовые, числовые и т. д. Например, в результате анализа фразы «В 2017 году» будут выделены три токена: «В» – текстовый, «году» – текстовый , «2017» – числовой. Такие токены можно назвать простыми. Кроме этого, выделяются метатокены – сложные токены, которые объединяют несколько простых токенов, например, существительные с определителями, скобки, кавычки, ...

В системе существует пополняемый статический словарь терминов. В нее можно добавлять термины и затем проверять их наличие в тексте. Кроме этого, в системе можно формировать динамически подобные словари на основе анализа текста.

При анализе текста создается аналитический контейнер, в который помещаются выделяемые сущности, токены в определенной последовательности, статистика и т.д.

Блок морфологического анализа позволяет определить класс словоформы (существительное, глагол, прилагательное и т. д.), род, падеж, число, язык (система поддерживает работу с тремя языками – русским, английским и украинским). В процессе обработки определяется, есть ли словоформа в словаре, отыскивается вариант словоформы в именительном падеже, в единственном числе; также анализируется дополнительная информация.

Программа PullEnti также способна выделять из сборников статей название статьи, авторов, текст статьи, список литературы.

Области применения программы PullEnti

Существует огромное количество различных предметных областей, которые описываются неструктурированной информацией. При этом, важно принимать качественные и своевременные решения, перерабатывая огромное количество информации разного типа, в том числе и текстовой. Можно привести примеры подобных областей, а именно, анализ резюме принимаемых на работу сотрудников, переработка сообщений СМИ, просмотр и сравнение информационно-рекламных материалов, почтовых сообщений, разбор сводок происшествий, справок по уголовным делам, архивных материалов и др.

Извлекаемая из текста информация должна быть адресной, потому из одного и того же текста можно извлекать совершенно различные виды информации, характерные для конкретной предметной области. В результате анализа текстовой информации выделяются типизированные объекты предметной области.

Программа PullEnti стала основой для построения множества систем: программа «Доктор Ватсон», система поиска экспертов, процессор BREF и другие.

Программа «**Доктор Ватсон**» предназначена для исследования массивов текстовой информации с целью выявления сущностей и связей между ними. В процессе работы загружаются тексты и, в результате анализа, выводится список сущностей и связей. При этом, пользователь может добавить недостающие сущности и связи (которые не были выделены программой), настроить выдаваемую информацию, сформировать отчет о результатах работы программы. Данная программа может использоваться в таких сферах деятельности, как криминалистика, конкурентная борьба, маркетинг, реклама, безопасность, разведка системными аналитиками.

Результат работы программы – отчет об исследуемом объекте, диаграммы сущностей и связей, который представлен на рис. 1. Из текущего текста выделены организации, персоны, их связи:

На рисунке 2 представлены выделенные объекты, их связи. Для каждой связи выделяется тип связи и название, например (например, Тип связи – «родственные», Заголовок связи – «отец»; Тип связи – «владение», Заголовок связи – «особняк в центре Вашингтона» и т. д.) определяются попарно объекты-участники связи. Для более полного определения ситуации выделяется не только время, характеризующее текущую ситуацию, но и интервалы времени. Дополнительные параметры позволяют выяснить, является ли связь симметричной для данной пары выделенных объектов (субъектов). Также для каждого выделенного объекта (субъекта) выделяются атрибуты. Например, для типа объекта "персона"выделяются имя, фамилия, отчество, дата рождения и т. д.

Более подробные данные, в случае, если они описаны в анализируемом тексте, выводятся в разделе «Отчеты» (вкладка «Отчеты»), в который попадает информация не только о фамилии, имени и отчестве персоны, но и возраст, дата рождения, паспортные данные, место проживания, контактные данные, данные об образовании и сроки обучения.

Результаты работы программы могут быть представлены в виде графа (вкладка «Диаграммы»), см. Рис. 3. В отчете выводятся обнаруженные объекты (персоны, организации, локации, атрибуты, ...), их связи в виде, удобном для анализа.

Программа «Логика ECM. Правовая экспертиза» предназначена для автоматизации процесса проведения экспертизы проектов нормативно-правовых актов (НПА), организационно-распорядительных документов, договоров и других документов. Система значительно упрощает процесс проведения

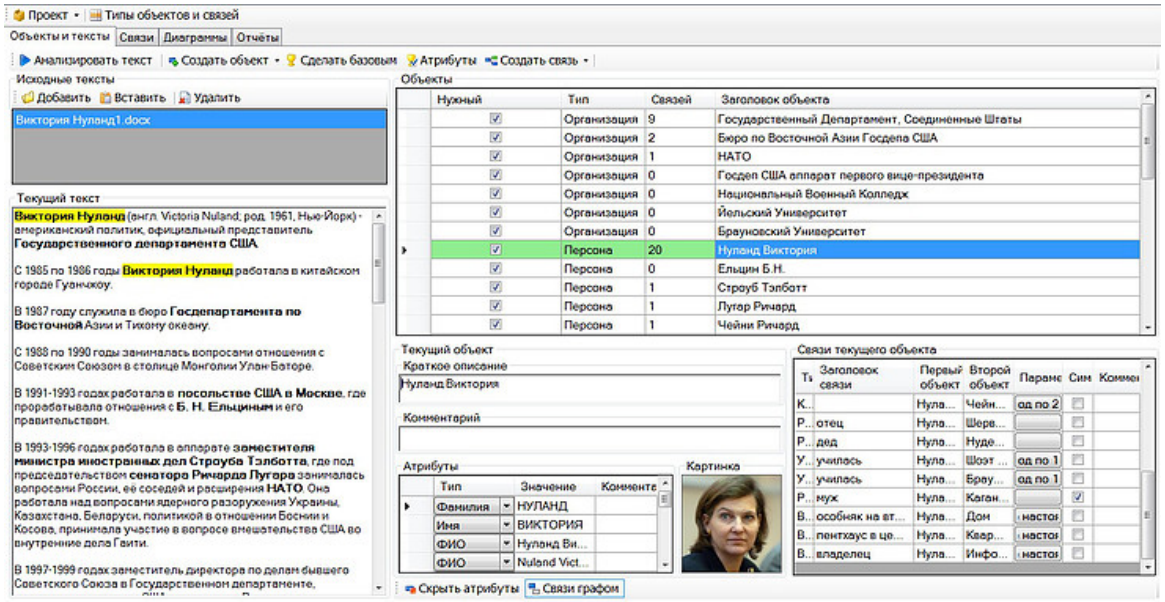

Рис. 1: Результаты работы программ «Доктор Ватсон». Выделение сущностей.

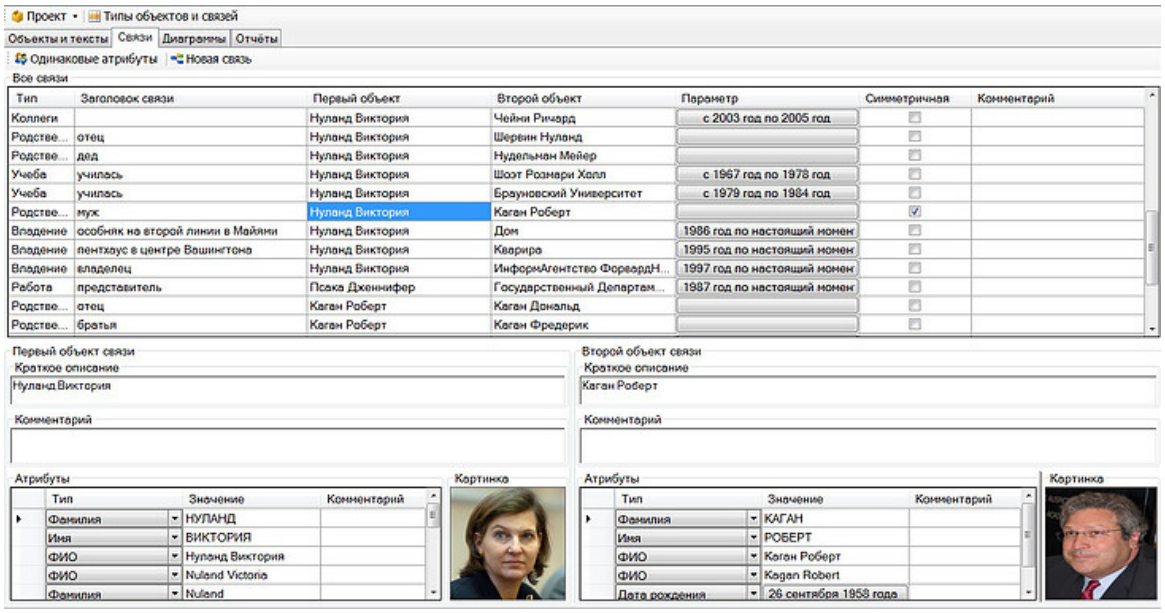

Рис. 2: Выделение связей, периодов.

правовой экспертизы и сокращает его сроки, выполняя рутинные операции и кардинально снижая затраты рабочего времени квалифицированных юристов.

«Логика ECM. Правовая экспертиза» автоматически, за несколько секунд поможет, например, установить:

— Не содержатся ли в проверяемом документе ссылки на нормативные правовые акты, которые утратили силу?
— Нет ли в проверяемом документе фрагментов других документов, не возникает ли избыточное дублирование нормативной документации?
— Соответствует ли оформление и структура документа установленным в организации правилам?
— Нет ли ошибок в оформлении цифровой информации в договоре, соответствуют ли друг другу суммы, указанные цифрами и прописью, правильно ли рассчитан НДС и т. п.

Система поиска экспертов индексирует информационные потоки, с которыми работают эксперты, запоминаются взаимосвязи между сотрудниками, терминами, выражениями, словосочетаниями, понятиями, ... Определяется мощность связей, т. е. насколько часто тот или иной сотрудник исполь-

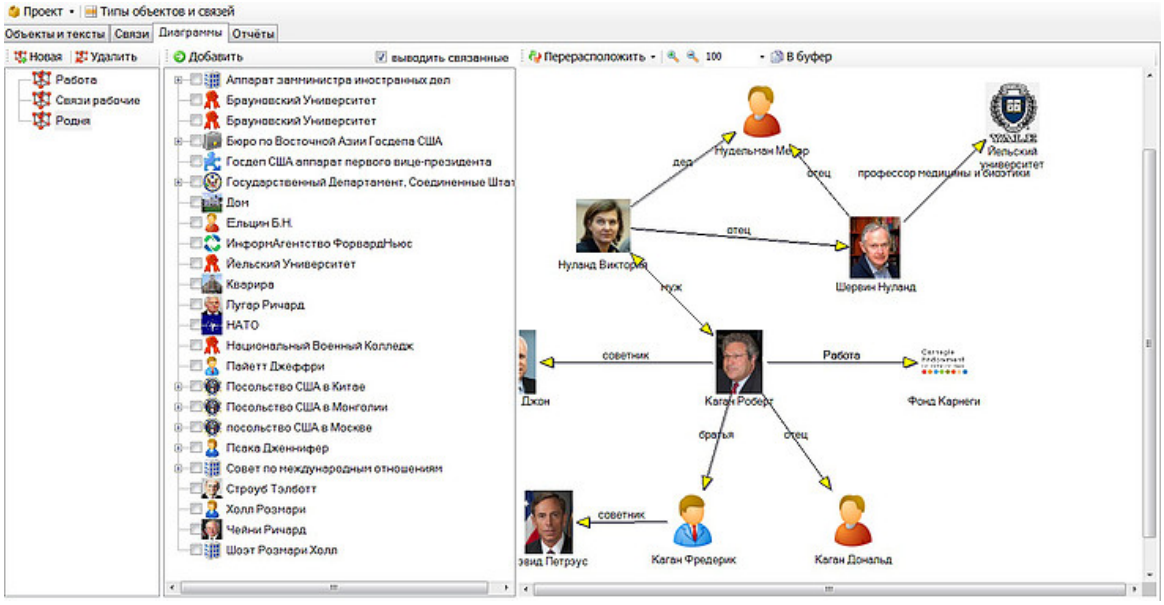

Рис. 3: Графическое представление результатов программы «Доктор Ватсон».

зует термины и понятия, насколько велико количество людей, работающих с термином. В результате в организации появляется возможность наиболее оптимального использования как экспертов, так и генерируемых ими идей. По тому, кто наиболее часто использует определенные термины можно быстро найти нужного эксперта.

Система сравнения диссертаций и поиск плагиата. Система ищет не только точно совпадающие фрагменты, но и фрагменты с перефразами (то есть с разнообразными заменами) Имеется сервис анализа (доступен после гуманоидной проверки) связей диссертантов – можно загрузить информацию о диссертациях, авторах, научных консультантах и оппонентах для поиска взаимосвязей (фабрик диссертаций). Диссертацию можно сравнивать саму с собой.

Процессор BREF

На основе лингвистического процессора Pullenti был реализован процессор обработки ссылок и списка литературы BREF, который позволяет по выделенной информации построить Граф Цитирования и Граф Соавторов, отражающие формальные связи в коллекции документов. Лингвистический Процессор (ЛП) также предназначен для выявления похожих по смыслу фраз в других статьях и в документах из Интернета. Лингвистический процессор настраивается с помощью метода машинного обучения.

Тексты статей из БД были пропущены через ЛП, с помощью которого выявлены названия статей, авторы и библиографические ссылки, а также упоминания авторов в текстах статей. Составлен рейтинг упоминаний авторов (РУА). Лингвистический процессор (ЛП) получает на вход полный текст статьи, а на выходе строит формальную структуру на языке XML следующего вида:

```
<RESULT>
    <ARTICLE>
    <AUTHOR>описание автора</AUTHOR>...
    // может быть несколько или не быть совсем
    <TITLE>наименование</TITLE>
    <YEAR>год, если есть</YEAR>
    <LANG>базовый язык: RU\EN </LANG>
    <LINK>
    <NUMBER>номер в списке литературы</NUMBER>
    <AUTHOR>...
    <TITLE>...
    <YEAR>...
    </LINK>
    <LINK> ... </LINK>
    ...
    </ARTICLE>
    <ARTICLE>
    ...
    </ARTICLE>
</RESULT>
```

В данной структуре формальные библиографические ссылки помечены метками `<LINK> ...</LINK>`. При описании автора `<AUTHOR>` в таком файле: LAST-фамилия, FIRST-имя, MIDDLE-отчество. Из подобных XML-файлов строится, например, граф цитирования.

С помощью программы BREF были обработаны сборники статей. В результате обработки сформированы XML-структуры. Ниже представлен фраг-

мент такой структуры, построенной на основе анализа сборника научных статей:

```
<?xml version="1.0" encoding="utf-8"?>
<result>
 <ARTICLE>
  <TITLE>
    О ВОССТАНОВЛЕНИИ ИЗОБРАЖЕНИЙ ПО КОДАМ
    В НЕКОТОРЫХ ВЫРОЖДЕННЫХ СЛУЧАЯХ
  </TITLE>
  <NORMALTITLE>
    О ВОССТАНОВЛЕНИИ ИЗОБРАЖЕНИЙ ПО КОДАМ
    В НЕКОТОРЫХ ВЫРОЖДЕННЫХ СЛУЧАЯХ
  </NORMALTITLE>
  <AUTHOR LAST="Агниашвили" FIRST="П"
    MIDDLE="Г" />
  <REF LANG="RU" YEAR="2001">
   <TITLE>
     ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ
     ЗРИТЕЛЬНОГО ВОСПРИЯТИЯ
   </TITLE>
   <NORMALTITLE>
     ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ ТЕОРИИ
     ЗРИТЕЛЬНОГО ВОСПРИЯТИЯ
   </NORMALTITLE>
   <AUTHOR LAST="Козлов" FIRST="Вадим"
     MIDDLE="Никитович" />
   <SOURCE>[1] Козлов В.Н. Элементы
     математической теории зрительного
     вос- приятия. - М.: Изд-во Центра
     прикладных исследований при механико-
     математическом факультете МГУ, 2001.
   </SOURCE>
  </REF>
  <REF LANG="RU" YEAR="2005">
   <TITLE>НАЧАЛА АЛГЕБРЫ: ЧАСТЬ I</TITLE>
   <NORMALTITLE>
     НАЧАЛА АЛГЕБРЫ ЧАСТЬ I
   </NORMALTITLE>
   <AUTHOR LAST="Михалев" FIRST="А"
     MIDDLE="А" />
   <AUTHOR LAST="Михалев" FIRST="А"
     MIDDLE="В" />
   <SOURCE>[2] Михалев А. А., Михалев А. В.
     Начала алгебры: Часть I. - М.:
     Интернет-университет информационных
     технологий, 2005.
   </SOURCE>
  </REF>
  <REF LANG="RU" YEAR="1987">
   <TITLE>С ТОЧКИ ЗРЕНИЯ ВЫСШЕЙ</TITLE>
   <NORMALTITLE>
     С ТОЧКИ ЗРЕНИЯ ВЫСШЕЙ
   </NORMALTITLE>
   <AUTHOR LAST="Клейн" FIRST="Ф" />
   <SOURCE>[3] Клейн Ф. Элементарная
     математика с точки зрения высшей:
     Том 2. Геометрия. - М.: Наука, 1987.
   </SOURCE>
  </REF>
 </ARTICLE>
```

Данная программа позволяет выделять наиболее значимые научные направления, наиболее цитируемые идеи и их авторов. С ее помощью можно строить индекс цитируемости научных статей.

Особенности анализа программой BREF текстов статей в сборниках научных трудов

Тестирование программы BREF проходило на достаточно представительном массиве документов. Были проанализированы несколько сборников научных трудов.

В процессе отладки программы BREF возникало большое количество нестандартных ситуаций, которые приводили к ошибкам. Были выделены следующие типы ошибок распознавания:

— неправильно распознается название статьи;
— неправильно выделен автор статьи;
— пропущен заголовок статьи, автор, источник в списке литературы;
— допущены ошибки в формировании XML-структуры документа по причине некорректного извлечения информации и текста;
— не определяется источник в списке литературы.

Сложности выделения названия статьи возникают по причине наличия нарушений в структуре статьи. Например, заголовок статьи может располагаться на одной строке с фамилией автора, заголовок располагается на нескольких строках, список предыдущей статьи завершается или начинается некорректно и т. д. В случае, если фамилия автора не отделена от заголовка пробелом, используется подход выделения фамилий из текста с учетом стандартных окончаний.

Выделенные ошибки позволили в существенной степени подкорректировать работу программы, сделать ее работу более устойчивой.

Очень часто подобные ошибки являются результатом неправильного оформления документов. И программа выявляет подобные ситуации. Соответственно, ее можно использовать для контроля структуры документов при подготовке к опубликованию в печати или в сети Интернет.

Результаты соревнования по выделению сущностей

В этом разделе представлены результаты соревнования по выделению сущностей и извлечению фактов на международной конференции по компьютерной лингвистике «Диалог-2016».

Процессом распознавания текстов естественного языка занимаются достаточно давно во всем мире. Ввиду сложности проблемы извлечения структурированной информации из текстов на этом пути остается еще много нерешенных задач. Особенно сложно эта задача решается для текстов русского языка, как одного из наименее структурированных и сложных языков. Во многих странах проводятся соревнования между разработчиками систем по извлечению информации из текстов естественного языка.

В рамках проводимого соревнования Dialogue Evaluation в Москве было проведено соревнование по извлечению информации из текстов естественного русского языка.

Лингвистический процессор PullEnti под псевдонимом Pink на соревновании FactRuEval конференции Диалог-2016 занял первые места на большинстве дорожек
(см. http://www.dialog-21.ru/evaluation/2016/ner/).
Соревнование проводилось на следующих дорожках:

— определение в тексте границ именованных сущностей, таких как персона, организация, локация;
— выделение именованных сущностей с определением атрибутов в нормализованном виде. Для персон это фамилия, имя и отчество. Для организаций и локаций – нормализованное название;
— извлечение фактов (например: «встреча», «покупка», ...) и наборов строковых полей (например: «участник встречи 1», «участник встречи 2», «место встречи», «дата/время начала встречи», ...);

В результате программа PullEnti (псевдоним Pink) победила в соревновании по выделению именованных сущностей и их атрибутов.

Ближайшие планы по доработке программы PullEnti

Следующими шагами исследования станут:

— Уточнение границ моделируемой предметной области за счет построения ядра ключевых фраз (в т. ч. с использованием методов вероятностного тематического моделирования [5, 7]);
— Выделение массива неявных ссылок (упоминаний авторов и идей, выраженных ключевыми фразами/значимыми словосочетаниями);
— Расчет корреляции между явными и неявными ссылками в рамках созданной коллекции.

Заключение

Разработанная программа PullEnti может быть использована в различных областях, в которых информация представлена в текстовом виде. Особенно это важно в тех случаях, когда необходимо выделять важную информацию из большого потока документов естественного языка. Очень хорошо работает данная технология в задачах кластеризации текстов по определенным признакам. При этом, существует возможность автоматической настройки программы на требования пользователя.

Описанные выше программы, созданные на основе технологии PullEnti, доказывают ее эффективность в самых различных областях.

Литература

[1] Золотарев О.В., Козеренко Е.Б., Шарнин М.М. Проведение аналитической разведки на основе анализа неструктурированной информации из различных источников, включая интернет и средства массовой информации // Вестник Российского нового университета. - М.: РосНОУ, 2015. № 9, с. 49-54.

[2] Шарнин М.М., Шагаев И., Протасов В.И., Родина И.В., Золотарев О.В., Попова О.А., "Использование веб-семантики для совершенствования образовательных программ вузов Вестник МГГУ им. М.А.Шолохова, Филологические науки, 2015, № 2, с.97-112.

[3] М.М. Шарнин, О.В. Золотарев, Н.В. Сомин. Извлечение и обработка знаний из неструктурированных текстов деловой сферы и социальных сетей, 4-я международная научно-практическая конференция "Социальный компьютинг: основы, технологии развития, социально-гуманитарные эффекты Москва, МПГУ, 2014, 22-24 октября.

[4] Oleg V. Zolotarev, Michael M. Charnine, Andrei G. Matskevich, Konstantin I. Kuznetsov. Business Intelligence Processing on the Base of Unstructured Information Analysis from Different Sources Including Mass Media and Internet. Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), vol.I, WORLDCOMP'15, July 27-30, 2015. Las Vegas Nevada, USA, v.I, pp.295-299.

[5] Irina V. Galina1, Michael M. Charnine, Nikolai V. Somin, Vladimir G. Nikolaev, Yulia I. Morozova, Oleg V. Zolotarev. Method for Generating Subject Area Associative Portraits: different Examples. Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015), vol.I, WORLDCOMP'15, July 27-30, 2015. Las Vegas Nevada, USA, v.I, pp.288-294.

[6] O. Zolotarev, M. Charnine, A. Matskevich. A Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts". Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014), vol.I, WORLDCOMP'14, July 21-24, 2014. Las Vegas Nevada, USA. CSREA Press, pp.82-87.

[7] Михеев М.Ю., Сомин Н.В., Галина И.В., Золотарев О.В., Козеренко Е.Б., Морозова Ю.И., Шар-

нин М.М. Фальштексты: классификация и методы опознания текстовых имитаций и документов с подменой авторства // Информатика и ее применения. Том 8, выпуск 4, РАН, - М.:, 2014.

[8] Золотарев О.В., Шарнин М.М. Методы извлечения знаний из текстов естественного языка и построение моделей бизнес-процессов на основе выделения процессов, объектов, их связей и характеристик. В сборнике: Труды Международной научной конференции СРТ2014 Международная научная конференция Московского физико-технического института (государственного университета) Института физико-технической информатики. Институт физико-технической информатики. 2015. С. 92-98.

[9] Золотарев О.В. Козеренко Е.Б., Шарнин М.М. Принципы построения моделей бизнес-процессов предметной области на основе обработки текстов естественного языка // Вестник Российского нового университета. - М.: РосНОУ, 2014. № 4. С. 82-88.

[10] Золотарев О.В. Методы выделения процессов, объектов, отношений из текстов естественного языка. // Проблемы безопасности российского общества. - Смоленск: Свиток, 2014, № 3-4, с. 276-283.

[11] Золотарев О.В. Инновационные решения в формировании функциональной структуры предметной области // Вестник Российского нового университета. - М.: РосНОУ, 2013. № 4. С. 82-84.

[12] Золотарев О.В. Методы и инструменты моделирования предметной области. В сб. трудов по материалам конференции «Цивилизация знаний: Проблемы социальных коммуникаций» – М.: РосНОУ, 2012.

[13] Золотарев О.В. Новые подходы в построении функциональной структуры предметной области. В сб. трудов по материалам конференции «20 лет постсоветской России: кризисные явления и механизмы модернизации». – Екатеринбург: Гуманитарный университет, 2011.

[14] Золотарев О.В. Формализация знаний о предметной области на основе анализа естественно-языковых структур. В сб. трудов по материалам конференции «Цивилизация знаний: Проблема человека в науке XXI века». – М.: РосНОУ, 2011.

[15] Золотарев О.В. Средства анализа информации в системах, основанных на семантических сетях. В сб. трудов по материалам конференции «Цивилизация знаний: Проблемы модернизации России». – М.: РосНОУ, 2010.