Извлечение информации из текста в системе ИСИДА-Т♣

© Д.А.Кормалев, Е.П.Куршев, Е.А.Сулейманова, И.В.Трофимов

Исследовательский центр искусственного интеллекта ИПС РАН
dkormalev@acm.org, epk@epk.botik.ru, yes@helen.botik.ru, itrofimov@km.ru

Аннотация

Статья посвящена методам и подходам к извлечению информации из текста на естественном языке, реализованным в системе ИСИДА-Т. Основной акцент сделан на представлении знаний и распознавании текстовых ситуаций.

Введение

Значительная доля информации, доступной в электронном виде, представлена в виде текстов на естественном языке. Заключенная в них полезная информация не структурирована, а значит, ее невозможно обработать и проанализировать классическими вычислительными методами и средствами. Тексты могут быть прочитаны и поняты человеком, но для вычислительной машины они — всего лишь цепочки символов. Меж тем, машинная обработка информации существенно ускоряет любой рабочий процесс и обеспечивает качество результата. Объем накопленной текстовой информации заставляет задуматься о средствах автоматической обработки текстов.

Технология извлечения информации из текстов на естественном языке (ТИИ) [8] — это технология обработки текста, которая позволяет автоматически «просматривать» относительно большой объем текстов, содержащих относительно небольшое количество искомой информации. Обнаруженная в тексте информация преобразуется в структурированный формат: выявляются целевые факты, объекты, отношения в виде, пригодном для дальнейшей автоматической обработки (статистической обработки, визуализации, поиска закономерностей в данных и других).

Иногда ТИИ рассматривают как специфическую разновидность информационного поиска. Отличия ТИИ от информационного поиска заключаются в том, что «запросы» должны быть известны заранее, а результатом является не набор ссылок на документы, а построенные структуры данных, описывающие релевантные факты из набора документов.

Приведем несколько областей применения ТИИ:

- расширение возможностей информационного поиска (поиск не по ключевым словам, а по фактам, ситуациям, объектам, отношениям);
- построение досье на персон или организации из открытых текстовых источников;
- мониторинг сообщений СМИ (примеры событий, которые могут представлять интерес: слияния и поглощения компаний, появление новых игроков на рынке, выпуск новой продукции, теракты);
- извлечение специфической метаинформации из коллекций документов большого объема (например, построение по текстовой базе муниципальных нормативно-правовых актов, связанных с недвижимостью, реляционной базы данных с информацией о типах событий, объектах и субъектах).

Первоначально задача ТИИ формулировалась как выделение фрагментов текста, содержащих релевантную информацию, и, возможно, преобразование их в реляционную форму. Для решения задачи в такой постановке часто достаточно анализировать локальный контекст, используя ограниченный набор знаний предметной области. Назовем такую технологию *извлечением информации в «слабом» смысле*. Результаты извлечения информации в «слабом» смысле и характер их представления несколько ограничивают возможности дальнейшего использования добытых из текста данных. *Извлечением информации в «сильном» смысле* мы назвали бы переход от базы текстовых фактов к такому их представлению, которое можно было бы использовать как интеллектуальный информационный ресурс, своего рода базу текстовых знаний.

Наши исследования были направлены на усовершенствование методов и расширение возможностей ТИИ, что позволило бы подойти вплотную к решению задачи извлечения информации в «сильном» смысле. Полигоном для экспериментальной проверки идей и практического воплощения разработанных подходов стала система ИСИДА-Т[1], над которой мы работаем в течение нескольких лет.

Чтобы получить информацию из прочитанного фрагмента текста (понять текст), человек должен знать язык, на котором написан текст, и располагать некоторым объемом «фоновых» знаний. Аналогично, система извлечения информации из

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

текста должна располагать двумя видами средств: средствами анализа естественного языка и некоторым объемом знаний предметной области. Однако прежде чем приступать к рассмотрению этих средств, остановимся на общей организации и инфраструктуре системы.

1 Общая организация системы

Краеугольным камнем системы ИСИДА-Т является точная настройка на предметную область и конкретную задачу извлечения. С одной стороны, это достигается за счет редактирования лингвистических ресурсов, ресурсов знаний, правил извлечения и правил трансформации. С другой стороны, настройка может потребовать включения в процесс обработки дополнительных специализированных методов обработки текста. Кроме того, для каждой задачи необходимо подобрать наиболее подходящие алгоритмические средства анализа из набора имеющихся. Эти аспекты требуют создания такой архитектуры, при которой легко могут добавляться и замещаться алгоритмические компоненты процесса извлечения.

Проблема конфигурирования на алгоритмическом уровне потребовала создания модульной архитектуры и декларативного подхода к определению процесса извлечения. Модули получили название обрабатывающих ресурсов в противовес лингвистическим ресурсам и ресурсам знаний. В конфигурации декларируется порядок обработки документа аналитическими модулями, потоки данных между ними, а также параметры их работы.

Обрабатывающие ресурсы можно разделить на следующие группы.

- *Ресурсы предобработки.* Сюда относятся средства определения кодировки документа, извлечения текста и стилевой разметки из документа, предварительной фильтрации.
- *Ресурсы лингвистического анализа.* Осуществляют разбор текста на отдельные слова, морфологический анализ (в том числе специализированные варианты для различных категорий имен собственных), поверхностный синтаксический анализ и определение границ предложений.
- *Ресурсы извлечения.* Осуществляют поиск в документе целевой лексики и синтаксических конструкций, а также первичное структурирование информации.
- *Ресурсы унификации знаний и вывода.* Осуществляют унификацию и отождествление элементов знаний, вывод производных знаний.
- *Ресурсы подготовки результата.* Осуществляют приведение извлеченной информации к определенному формату и передачу за пределы последовательности обработки (в БД, глобальный ресурс знаний, файл, приложение).

В целом средства конфигурирования выполняют те же функции, что и каркас (framework) в известных системах GATE [10] и UIMA [11]. Эти средства обеспечивают (1) расширяемость архитектуры, (2) управление потоками данных, (3) поддержку моделей разделяемой памяти, (4) настройку параметров обрабатывающих ресурсов и последовательности обработки, (5) упрощение и унификацию процесса разработки новых обрабатывающих ресурсов. Проблемой хранения результатов анализа в нашем подходе занимаются, преимущественно, сами обрабатывающие ресурсы. Обрабатывающие ресурсы мы реализуем в формате динамически загружаемых библиотек (или разделяемых объектов).

2 Средства анализа естественного языка

Средства анализа естественного языка, используемые в ТИИ, можно разделить на две большие категории: средства общего лингвистического анализа и предметно-ориентированные методы распознавания текстовых ситуаций.

Средства общего лингвистического анализа включают в себя графематический, морфологический и синтаксический анализ. Эти средства применимы практически во всех предметных областях, существует ряд реализаций с довольно высокими показателями качества, поэтому мы не будем останавливаться на них подробно.

Вторая категория — средства и методы распознавания текстовых ситуаций, характерных для решаемой задачи и предметной области. Распознавание текстовых ситуаций состоит в выделении фрагментов текста, описывающих объекты, и содержательных связей между этими фрагментами, основанных в той или иной мере на синтаксисе естественного языка. Можно рассматривать распознавание ситуаций как ориентированный на предметную область частичный, но точный синтактико-семантический анализ.

Распознавание опирается на сопоставление образцу, который задается при помощи правил на специализированном формальном языке. Правила определяют не только образец, но и действия, которые должны быть выполнены при успешном сопоставлении. Правила работают не с текстом как последовательностью символов, а со структурами, построенными «над» текстом и выражающими лингвистическую и предметную информацию о нем.

Для упрощения конфигурирования системы очень желательно, чтобы все модули использовали одинаковый способ представления информации о тексте (разметки текста). В системе ИСИДА-Т все модули, в том числе средства общего лингвистического анализа, используют структуры данных, описанные ниже.

2.1 Разметка текста и структуры данных

В различных системах обработки текста на естественном языке используется широкий спектр средств для представления лингвистической и предметно-ориентированной информации о тексте в целом или его фрагментах. Единого подхода к представлению разметки текста и информации о нем не существует.

В последнее десятилетие довольно широко используется способ представления информации о тексте, основанный на так называемых *аннотациях*, отличающийся простотой и высокой степенью универсальности [12]. На сегодняшний день многие системы обработки текста в той или иной степени используют идеи модели аннотаций. Эта модель используется и в нашем подходе.

Аннотация — объект, который приписывается фрагменту текста (например, слову, словосочетанию, предложению, ссылке на сущность предметной области и т.д.) и описывает свойства этого фрагмента. Аннотации разбиты на конечное множество классов. Каждый класс аннотаций описывает текст в определенном аспекте. Информация о фрагменте представлена значениями именованных атрибутов аннотации. Наборы классов и атрибутов аннотаций намеренно не специфицированы, чтобы можно было использовать произвольный набор обрабатывающих модулей и представлять необходимую лингвистическую и предметную информацию. Обмен данными между модулями тоже идет в терминах аннотаций: новые аннотации могут строиться на основании полученных на предыдущих этапах анализа.

Из способа представления информации с помощью аннотаций следует возможность разработки средств анализа текста, компоненты которых слабо связаны между собой. Не отражающееся на функциональных характеристиках сложной системы уменьшение числа зависимостей между ее составляющими облегчает ее понимание, разработку и поддержку. Слабая связность является существенным преимуществом, так как повышает возможность повторного использования компонентов и снижает риск критических сбоев, вызванных неправильным взаимодействием компонентов (например, из-за того, что в цепочке обработки какой-то компонент ошибочно не был зарегистрирован, или же частично нарушился порядок обработки).

Впрочем, базовая модель аннотаций не лишена недостатков. В частности, она не подразумевает средств проверки соответствия атрибутов и их значений. Атрибуты могут быть только атомарными. Отсутствуют возможности установления связей между отдельными аннотациями. Нет средств для контроля расположения границ аннотаций разных классов, в то время как для большинства классов аннотаций можно задать условия, описывающие их взаимное расположение. Например, аннотации, описывающие синтаксис предложения в терминах системы составляющих, не могут пересекаться – для них возможно только отношение строгого вхождения или совпадения.

В реализации системы ИСИДА-Т модель аннотаций была дополнена некоторыми полезными средствами. В частности, было снято ограничение на атомарность атрибутов и добавлена возможность устанавливать ссылки между аннотациями.

2.2 Язык правил распознавания текстовых ситуаций

Для распознавания текстовых ситуаций используется набор правил, описывающих характерные для конкретной задачи способы выражения ситуации в тексте. Эти правила задают образец для сопоставления и действия, которые должны быть произведены после успешного сопоставления. Качество работы (полнота и точность) ТИИ тесно связано с возможностями языка правил. Ряд современных систем извлечения информации (в том числе, система ИСИДА-Т) берут за основу различные диалекты языка CPSL [7]. Использование этого языка подразумевает разметку текста при помощи аннотаций.

Единицей трансляции языка правил является фаза. Правила, входящие в одну фазу, применяются в недетерминированном порядке. Результаты фазы — изменения, внесенные в набор аннотаций после работы правил — фиксируются после применения всех правил и становятся доступны в последующих фазах. Поэтому правило не может использовать результаты работы другого правила из этой же фазы. Можно рассматривать фазу как модуль для специфического анализа текста. Работа фаз может перемежаться применением произвольных обрабатывающих ресурсов.

Правило — основная единица языка. Правила представляются в виде «образец→ действие». Здесь «образец» — образец для поиска в терминах высказываний о взаимном расположении и значениях атрибутов аннотаций разных классов (левая часть правила); «действие» — набор действий, выполняемых при успешном сопоставлении (правая часть правила). По структуре левая часть правила во многом схожа с регулярным выражением, но существенное отличие состоит в том, что роль символов в правиле играют тесты. Тест представляет собой конъюнкцию высказываний (элементарных тестов) о значениях атрибутов аннотаций разных классов. Из тестов могут образовываться сложные конструкции с использованием следования, альтернативы, квантификаторов и скобок. Чтобы обозначить границы фрагментов текста, сопоставленных подвыражениям, используются метки. Метка — это идентификатор, которым помечается образец. В дальнейшем (при выполнении действий в правой части правила) можно использовать метку для ссылки на фрагмент текста, сопоставленный подвыражению.

Язык правил, используемый в системе ИСИДА-Т, является расширением CPSL. Предлагаемые нами расширения преследуют две цели: 1) обеспечить возможность описывать более сложные контексты, в которых встречается целевая информация, и 2) снизить объем рутинной работы при создании системы правил за счет более компактного описания контекста [5].

Отличия от других реализаций, например, JAPE [10] или диалекта CPSL, используемого в продуктах RCO [3] состоят в следующем.

- Для передачи информации между элементарными тестами, а также в правую часть правил могут использоваться именованные переменные, значения которых присваиваются явно в ходе сопоставления. Множество значений переменных входит в контекст сопоставления. Использование переменных позволяет компактно описывать отношения между атрибутами аннотаций, рассматриваемых в разных элементарных тестах. В частности, этот механизм обеспечивает компактное описание согласования языковых единиц, рассматриваемых в различных состояниях конечного автомата.
- Реализована встроенная поддержка расширенного спектра типов данных, в том числе, ссылок на аннотации и множественных значений. Данные этих типов могут использоваться в качестве значений переменных и значений атрибутов аннотаций.
- Логика работы интерпретатора правил приведена в максимальное соответствие поведению интерпретатора обычных регулярных выражений. Отличия от современной реализации JAPE и Montreal transducer [14] заключаются в поддержке «жадных» и «нежадных» квантификаторов и опережающей проверки.
- Поддерживаются кванторы существования (по умолчанию) и всеобщности, связывающие элементарные тесты. К кванторам может добавляться отрицание.
- Существуют языковые средства, позволяющие гибко проверять взаимное расположение аннотаций, рассматриваемых в контексте сопоставления, и прочих аннотаций во входной коллекции.
- В тестах могут использоваться функции для обращения к ресурсу знаний (раздел 3), например, проверки таксономической принадлежности элементов. Для более сложных запросов к ресурсу знаний используется предметно-ориентированный язык, совпадающий с языком описания левой части правил трансформации (подраздел 3.2).

Общая проблема средств распознавания текстовых ситуаций — при расширении функциональных возможностей этих средств резко падает производительность. Для решения этой проблемы мы использовали два основных способа оптимизации интерпретатора правил: предобрабатывать правила, анализируя потоки управления [9, 13], и сокращать перебор кандидатов при выполнении тестов [4]. Внедрение этих модификаций позволило ускорить интерпретацию правил в среднем в 6 раз в зависимости от конфигурации системы и качества входных данных (в отдельных случаях наблюдался прирост производительности до двух порядков). В большинстве случаев повышение производительности сопровождалось снижением расхода памяти на 20-40%.

3 Ресурс знаний

Практически в любой предметной области для точного извлечения требуются априорные знания о ней — знания о понятиях, объектах и отношениях, связанных с целями извлечения или являющихся целями. В свою очередь, извлеченная из текстов информация может нести в себе новые знания о предметной области и быть полезна для дальнейшей автоматической обработки текста. Тесная связь между априорной и извлеченной информацией, а также между предметными и лингвистическими знаниями сформировала потребность в унификации средств представления.

3.1 Представление знаний

Интегрированный ресурс знаний (РЗ) [1] системы ИСИДА-Т объединяет в себе базу априорных предметных знаний, хранилище фактографической информации и словарь. Предметные знания хранятся в РЗ в структурах, называемых *элементами знаний*. Элементы знаний делятся на 4 категории [6]: 1) концепты (CT), 2) экземпляры концептов (CI), 3) типы предметных отношений (RT), 4) экземпляры отношений (RI). Наш подход к представлению знаний использует элементы семантических сетей и систем фреймов.

Концепты и типы отношений служат для представления онтологической информации о предметной области и задаются априорно. Экземпляры концептов и отношений составляют базу фактов предметной области и могут быть как априорными, так и извлеченными из текстов.

Для каждого элемента знаний задается набор атрибутов. В списках атрибутов CT и RT хранятся пары «имя—ограничения на значение», в списках атрибутов CI и RI — пары «имя—значение». В терминах системы фреймов CT и RT выражались бы прототипами фреймов, а CI и RI — экзофреймами. Неявно определены два специальных (служебных) типа отношений: ISA и AKO. Их интерпретация такая же, как в системах фреймов.

Лингвистическая составляющая ресурса знаний — словарь. Словарь связан с базой предметных знаний посредством ссылок от дескрипторов к элементам знаний: дескрипторы словаря базовой лексики ссылаются на концепты, а дескрипторы словаря собственных имен — на априори известные экземпляры концептов из базы фактов. В отличие от тезауруса, дескрипторы в словаре базовой предметной лексики не связаны друг с другом никакими парадигматическими отношениями (последние выражаются с помощью отношений между соответствующими элементами базы предметных знаний).

Словарь предоставляет возможность указывать дополнительные ограничения на все словоформы, входящие в состав дескриптора и синонимов, чтобы увеличить точность распознавания словарных единиц в тексте.

Унификация априорных и извлеченных из текстов знаний удобна тем, что позволяет использовать одни и те же средства для работы с обоими типами знаний. Объединение лингвистических и предметных знаний в одном ресурсе, во-первых, облегчает первичное наполнение и последующую поддержку, а во-вторых, дает возможность использовать предметные знания уже на этапе первичной обработки текста правилами извлечения информации. Благодаря специально разработанному языку запросов к РЗ правила могут не ограничиваться словарной информацией, а обращаться в онтологию и базу фактов для проверки различных условий, требующих навигации по отношениям.

3.2 Трансформации

После извлечения информации из текста и помещения ее в хранилище фактографической информации часто требуется дополнительная обработка для ее унификации и уточнения. На основе такой обработки может решаться целый спектр задач:

- навигация по связанным объектам, фактам и ситуациям;
- определение и объединение тождественных элементов (некоторые случаи разрешения кореферентности);
- кластеризация сходных сюжетов;
- вывод имплицитной фактографической информации;
- генерация текстовых описаний фрагментов фактографической базы.

Для проведения экспериментов по преобразованию извлеченной фактографической информации был разработан язык трансформаций и выполнена экспериментальная программная реализация интерпретатора этого языка.

Трансформацию элементов ресурса знаний можно рассматривать как особый вид немонотонного вывода на знаниях. При трансформации происходит поиск образца ситуации в ресурсе знаний и выполнение указанных действий. Для описания ситуации можно задавать ограничения на типы элементов знаний, их атрибуты, наличие или отсутствие отношений того или иного типа между ними. Попытка выполнить действия производится для каждого набора элементов знаний, для которых выполняются условия, указанные в посылке правила трансформации. Набор действий включает в себя создание, удаление, модификацию элементов знаний, манипулирование их атрибутами.

Особенностью языка правил трансформации является сочетание декларативных и императивных элементов.

Язык трансформаций предназначен для описания правил, по которым выполняется преобразование элементов в хранилище фактографической информации. Правила языка схожи с продукционными правилами: каждое правило содержит образец для поиска (левая часть правила) и набор действий (правая часть), которые необходимо выполнить, когда образец был обнаружен.

При поиске образца используются следующие элементарные условия (перечислены только основные):

- проверка принадлежности элемента знаний к указанному классу;
- проверка наличия или отсутствия отношения указанного класса между элементами знаний;
- сравнение ссылок на элементы знаний;
- сравнение значений атрибутов элементов знаний.

В условиях левой части используются два вида переменных: переборные и присваиваемые. Для переборных переменных выполняется перебор возможных значений с означиванием переменных. Значения присваиваемых переменных устанавливаются явным образом, например, как результат функции или значение атрибута элемента знаний. Набор означенных переборных переменных и установленных присваиваемых переменных определяет контекст применения правила.

Все условия в левой части правила связаны конъюнкцией, соответственно, для выполнения условий правила должны выполниться все элементарные условия. После успешного означивания переборных переменных и выполнения всех элементарных условий происходит сохранение контекста применения правил для использования в правой части правила.

Правая часть правила представляет собой составной оператор простого императивного языка. В настоящее время реализованы следующие элементы:

- следование;
- составной оператор;
- условный оператор;
- присваивание значений переменным;

- вызов встроенных функций языка (создание, удаление, модификация, объединение элементов знаний).

Правила сгруппированы по фазам применения. Результаты применения правил и их побочные эффекты «незаметны» правилам, отнесенным к той же фазе — только правилам из последующих фаз. Если в результате ошибки записи правил или побочных эффектов других правил той же фазы выполнение всех действий правой части невозможно, происходит отмена эффекта частично выполненной правой части, после чего выполнение действий продолжается для других контекстов. Например, выполнение действий может быть невозможно, если элемент знаний, присутствующий в контексте одного правила, был удален в результате выполнения правой части другого.

Очевидно, что полный перебор всех возможных вариантов для означивания переменных в левой части правила неэффективен. Для повышения эффективности при поиске контекстов, в которых выполняется сопоставление, были разработаны алгоритмы предобработки правил трансформации и подготовки вспомогательных структур в ресурсе знаний, с которым будет идти работа. Цель предобработки — выделение минимально возможных множеств кандидатов для означивания переменных в правилах. Поскольку в пределах фазы (до начала выполнения действий) ресурс знаний не изменяется, можно однократно создать вспомогательные индексы и пользоваться ими при сопоставлении образцов всех правил, входящих в фазу. Индексы могут использоваться совместно всеми правилами фазы. С использованием индексов происходит исключение элементов знаний, которые заведомо не могут участвовать в означивании переменных.

После построения множеств кандидатов для означивания переменных происходит перебор кортежей декартова произведения множеств кандидатов для каждой переменной и окончательная проверка выполнения условий. Это необходимо, потому что не для всех условий возможна предобработка; кроме того, могут существовать зависимости между переборными переменными, которые можно определить только на этапе собственно сопоставления.

При успешном сопоставлении и означивании всех переборных переменных полный контекст сопоставления отправляется в хранилище результатов фазы. В дальнейшем это хранилище используется для выполнения действий каждого правила, входящего в фазу.

Правила трансформации позволяют унифицировать представление типовых ситуаций в хранилище фактографической информации и подготовить информацию для дальнейшей обработки, в том числе, для использования при анализе других текстов.

4 Результаты экспериментов

Чтобы читатель мог получить представление о качественных и технических характеристиках системы, рассмотрим задачу извлечения, которую мы решаем в настоящее время, и параметры системы, при которых эта задача решается.

Предметная область охватывает политику, межгосударственное взаимодействие и дипломатию; государственное и региональное управление; экономику, финансы, бизнес.

Целевые факты представляют собой события и состояния, участниками которых выступают целевые сущности.

К целевым сущностям относятся:

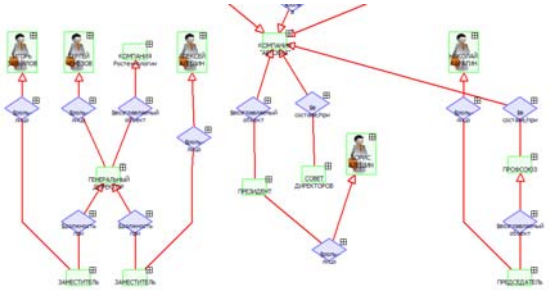
- лица;
- организации;
- роли лиц — должности, звания, род занятий, общие понятия принадлежности и иерархии (служащий, сотрудник, глава и т.п.), межличностные роли (родство, личное знакомство), членство и др.
- геополитические единицы.

Целевые факты описывают:

- отставки, назначения, пребывание в должности (роли);
- структурные отношения на множестве ролей лиц, организаций и геополитических единиц — должность при (лице, должности), должность в, должность во главе (организации или геополитической единицы), лицо во главе (организации или геополитической единицы), членство в организации, административно-территориальная принадлежность организации и др.;
- структурные отношения между организациями — в составе/при, часть-целое;
- отношения между лицами — родство (степень родства), знакомство и т.п.

Приведем характеристики системы на настоящий момент. В ресурсе знаний выделен 41 тип целевых и вспомогательных отношений (бинарных и тернарных) и 269 типов объектов предметной области (концептов), экземпляры которых могут стать участниками целевых ситуаций. Для решения задачи сейчас используется 156 контекстных правил извлечения информации, (42 фазы), а также 31 правило трансформации (6 фаз). Суммарная скорость обработки текста для такой конфигурации составляет порядка[2] 1 КБ/с на одном ядре процессора с тактовой частотой 2.4 ГГц.

Для примера на рисунке ниже приведен фрагмент результатов обработки новостной заметки [2].



Заключение

Описанные в методы и подходы могут найти применение в технологических цепочках хранилищ знаний, для построения и наполнения ресурсов знаний разного рода, для повышения точности и обогащения результатов работы поисковых машин. Методы обработки текста и работы со знаниями, реализованные в системе ИСИДА-Т, создают основу для средств извлечения информации в «сильном» смысле. Такие средства не ограничиваются разметкой текста; они подразумевают переход от корпуса текстов к такому представлению фактографической информации, которое можно было бы использовать как интеллектуальный информационный ресурс, своего рода базу текстовых знаний.

Литература

[1] Александровский Д.А., Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Модель и реализация ресурса знаний в системе извлечения информации из текста // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2008, 28 сентября–3 октября 2008 г., г. Дубна, Россия): Труды конференции. Т. 2. — М.: ЛЕНАНД, 2008. — С. 201-209.

[2] Госкорпорация "Ростехнологии" и АВТОВАЗ создадут холдинг по производству автокомпонентов, 2008. http://quote.rbc.ru/stocks/news/2008/06/27/31997446.shtml

[3] Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. – Москва, Наука, 2004.

[4] Кормалев Д.А. Повышение производительности при распознавании текстовых ситуаций // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2008, 28 сентября–3 октября 2008 г., г. Дубна, Россия): Труды конференции. Т. 2. — М.: ЛЕНАНД, 2008. — С. 192-200.

[5] Кормалев Д. А., Куршев Е. П. Развитие языка правил извлечения информации в системе ИСИДА-Т // Труды международной конференции «Программные системы: теория и приложения». — Т. 2. — М.: Физматлит, 2006. — С. 365-377.

[6] Сулейманова Е.А. Классификация ресурсов знаний в системе извлечения информации из текста // Математические методы распознавания образов: 13-я Всероссиийская конференция. Ленинградская обл., г. Зеленогорск, 30 сентября - 6 октября 2007 г.: Сборник докладов. — М.: МАКС Пресс, 2007. — С. 625—628.

[7] Appelt D.E. The Common Pattern Specification Language: Technical report / SRI International, Artificial Intelligence Center. — 1996.

[8] Appelt D. E., Israel D. J. Introduction to Information Extraction. Tutorial // Sixteenth Int. Joint Conf. on Artificial Intelligence IJCAI'99, Stockholm, Sweden, 1999.

[9] Cooper K. D., Harvey T. J., Kennedy K. A Simple, Fast Dominance Algorithm. Software Practice and Experience, 2001.

[10] Cunningham H., Maynard D., Bontcheva K., Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[11] Ferrucci D., Lally A.UIMA by Example. IBM Systems Journal 43, No. 3,. 455-475 (2004).

[12] Grishman R. TIPSTER Text Architecture Design. Version 3.1. — New York: NYU, 1998.

[13] Lengauer T., Tarjan R.E. A fast algorithm for finding dominators in a flow graph. ACM Transactions on Programming Languages and Systems, 1(1):115120, July 1979.

[14] Plamondon L. The Montreal Transducer module for GATE. http://www.iro.umontreal.ca/~plamondl/mtltransducer/1_1/README.html

Information extraction in ISIDA-T system

D.A. Kormalev, E.P. Kurshev, E.A. Suleimanova, I.V. Trofimov

The article discusses methods and techniques for information extraction from natural-language texts, as they are implemented in ISIDA-T system. Emphasis is made on knowledge representation and recognition of textual situations.

♣ Работа поддержана РФФИ, проект 09-07-00407, и программой фундаментальных исследований Президиума РАН №3, проект «Высокопроизводительные масштабируемые средства работы с фактографическими базами большого объема».

[1] Интеллектуальная система извлечения данных и их анализа (для обработки текстов).

[2] Скорость обработки текста и трансформаций сильно зависит от входного текста, а также объема накопленной фактографической информации.