УДК 519.767.6; 81’322.2

Е. А. Сидорова

Институт систем информатики им. А. П. Ершова СО РАН
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия
E-mail: lena@iis.nsk.su

ИНСТРУМЕНТАЛЬНЫЕ СРЕДСТВА ФАКТОГРАФИЧЕСКОГО АНАЛИЗА ДОКУМЕНТОВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ, ОСНОВАННЫХ НА ОНТОЛОГИЯХ

В статье рассматривается технология создания сервиса фактографического анализа текстовых ресурсов для различных информационных систем. Используемая база знаний включает четыре компонента: онтологию предметной области, словарь предметной лексики, модель рассматриваемых документов и схемы фактов, которые описывают структуру фактов и связывают термины словаря с элементами онтологии. Определяется набор инструментов, необходимых как для формирования базы знаний, так и для автоматического применения знаний эксперта при обработке текста.

Ключевые слова: анализ текста, лингвистическая онтология, извлечение фактов.

Введение

Современные требования к качеству автоматического анализа текстовых ресурсов приводят разработчиков к необходимости ограничивать как предметную область содержания документа, так и стиль оформления текста. Кроме того, большинство информационных систем, если они не являются специализированным литературным ресурсом, работают с документами, написанными в жанре «деловой прозы» [Ершов, 1994]. Однозначный контекст, строгое выражение мысли в таких текстах значительно уменьшают количество способов выражения одной и той же информации. Служебная либо общезначимая лексика, характерная для текста на естественном языке, исполняет роль связывания или замещения значимой лексики; не используются образные, литературные, эмоциональные выражения.

Основным инструментом, с помощью которого в настоящее время осуществляется информирование специалистов / пользователей, являются информационные системы (ИнС). В процессе развития ИнС сформировались три вида информационного обслуживания – документальное, фактографическое и концептографическое. Сущность документального обслуживания заключается в том, что пользователям предоставляют тексты документов, необходимые сведения из которых пользователь извлекает самостоятельно. Фактографическое обслуживание предполагает предоставлять информацию в виде сведений (отдельных данных, фактов, концепций). Эти сведения предварительно извлекаются из текстов документов и после определенной их обработки предоставляются пользователям. При концептографическом обслуживании документы и полученные сведения подвергаются интерпретации, оценке, обобщению. Возможности современных ИнС сводятся к фактографическому обслуживанию.

Большинство требований к сервису автоматического анализа текста, поддерживающему фактографическое обслуживание, сводится к задаче преобразования слабоструктурированного текста к хорошо структурированной информации [Рубашкин, 2006]. Основные отличия между такими сервисами заключаются в предметной области и, как следствие, структуре извлекаемых знаний. Таким образом, позволив пользователю настраивать предметные знания (в том числе и лингвистические) и разработав универсальный механизм, использующий эти знания для анализа документа, мы создадим технологию «быстрого» конструирования сервисов анализа для различных ИнС. Созданию такой технологии и посвящена данная работа. Она ориентирована на ИнС, которые используют явно выраженные (в виде онтологии) зна-

ния о предметной области [Боровикова, Загорулько, 2002]. Применение онтологий является одним из наиболее перспективных направлений исследований, поскольку позволяет формализовать и унифицировать операции обработки информации для повышения качества различных информационных услуг и сервисов. Онтологические знания, а также данные ИнС могут использоваться при автоматической обработке текста.

Роль онтологии при обработке текста

До сих пор задача анализа текста на естественном языке рассматривалась многими исследователями независимо от той обстановки, где ее результаты планировалось использовать. В отличие от работ, связанных с задачей полного извлечения смысла или извлечения всей информации из текста документа, для большинства ИнС нет необходимости делать полный семантический анализ всего связного текста. Были выделены следующие, полезные с точки зрения анализа, функциональные свойства онтологии ИнС.

Описание информационного объекта. Информационный объект (ИО) – это структурированная совокупность данных, представляющая описание некоторого объекта выбранной предметной или проблемной области. Каждый ИО соответствует некоторому классу онтологии (является экземпляром этого класса) и имеет заданную этим классом структуру. Между конкретными информационными объектами могут существовать связи, семантика которых определяется отношениями, заданными между соответствующими классами онтологии. Примерами ИО, помимо стандартных объектов предметной области, могут быть документы, видеоресурсы, интернет-сайты и др. Таким образом, никакой информационный объект или ресурс не может «присутствовать» в системе, если он не описан каким-либо понятием онтологии.

Описание содержания документа. Содержание текста документа описывается с помощью предметных ИО. Получившееся описание является контентом ИО, соответствующим документу [Васильев, Тузовский, 2003]. Создаваемые для использования в ИнС онтологии неполно охватывают содержание документов, они не являются семантической копией обрабатываемой информации, а отражают лишь те аспекты, которые существенны для решения конкретных задач в рамках системы.

Отражение схемы базы данных (БД). Данная функция позволяет обращаться к БД в терминах онтологии (например, поиск понятий, связанных с заданным, или объектов, являющихся экземплярами какого-либо понятия). Это свойство полезно при сопоставлении информации, полученной в результате анализа текста, и информации, уже присутствующей в системе.

Таким образом, онтология определяет формат данных, которые хранятся в ИнС, и, следовательно, определяет, какую именно информацию необходимо извлекать из текста документа, а какую можно проигнорировать. Результат анализа документа представляется в виде семантической сети объектов, являющихся экземплярами понятий и отношений, заданных онтологией предметной области.

Особенности деловой прозы

Прежде чем приступить к описанию той базы знаний, которая потребуется для обработки текста, необходимо выделить основные свойства объекта анализа (текста первичного документа) – жанра деловой позы [Ершов, 1994]. Для деловой прозы характерны следующие особенности.

Наличие строгой модельной ситуации, определяемой характером автоматизации или назначением создаваемой ИнС, для которой заданы правила распознавания и реакции на ее возникновение, хотя последовательность возникновения ситуаций может оставаться неопределенной. Это свойство приводит к тому, что деловая проза всегда внутренне формализована.

Ограниченность предметной области. Модель действительности определяется самой областью деловых отношений.

Ограничение естественного языка (т. е. используется концепция подъязыка как проекция общеупотребительного литературного языка на определенную предметную область и класс ситуаций общения). Потребность быстрого и точного взаимопонимания сделала язык деловой прозы четким, экономичным и жестким, а внешнее оформление текста документа –

структурированным. Поэтому мы можем вводить в систему соответствующие ограничения и делать больший упор на семантику текста, нежели на его синтаксическое представление.

Четкость функций каждого сообщения. Наличие цели, определяемой по заранее известным правилам, позволяет сконцентрировать анализ вокруг наиболее значимых понятий предметной области, к таким понятиям относятся, например, научный результат в научной статье или сообщение о какой-либо деятельности в деловом письме. Это свойство разительно отличает деловую прозу от других форм общения, например стихов или пространного и эмоционального повествования.

Лингвистическая онтология

Очевидно, что знаний о предметной и проблемной областях, хранящихся в онтологии, недостаточно для автоматического извлечения информации из текста – требуются дополнительные знания о языке, на котором эта информация представляется:

- словник или перечень минимальных единиц языка, используемых при описании значимой для ИнС информации;
- набор специфичных для заданного языка лингвистических знаний: морфологические классы, правила согласования терминов языка и т. п. (эти знания могут быть заданы с разной степенью подробности в зависимости от требований и возможностей разработчиков ИнС);
- знания о согласовании имеющихся лингвистических знаний с предметными знаниями, заданными онтологией ИнС (с этой целью термины группируются в семантические группы – семантические ориентации, которые в свою очередь также согласуются с элементами онтологии либо непосредственно, либо в соответствии с определенной схемой);
- набор описаний жанровых структур текста, соотнесенных с тем или иным типом текстовых ресурсов, хранящихся в ИнС.

В соответствии с данными требованиями была разработана лингвистическая онтология, содержащая всю необходимую для анализа информацию. Лингвистическая онтология – это четверка вида <O, V, F, D>, где O – онтология предметной и проблемной области ИнС; V – словарь терминов; F – множество упорядоченных наборов схем фактов (порядок отражает последовательность применения схем фактов во время анализа); D – множество моделей документов, для каждой из которых может быть определен собственный набор схем фактов.

Мы рассматриваем онтологию ИнС как часть лингвистической онтологии только для удобства формального описания. Каким образом связаны эти две онтологии на самом деле, требует дальнейшего осмысления. Рассмотрим подробнее компоненты лингвистической онтологии и их технологический аспект на примере предметной области «Хроники СО РАН».

Базовая задача анализа сообщений из архива хроник заключается в извлечении названий организаций – структурных подразделений РАН, упоминаний *персон*, их научных званий и ученых степеней, а также выявление связей между персонами и организациями. Таким образом, онтология ИнС включает:

```
class Персона (Фамилия: string; Имя: string; Отчество: string;
               Инициалы: string; ПолноеИмя: string; Звание: domen_Звания);
class Организация (Имя: string; Аббревиатура: string);
        class Институт : Организация;
        class Филиал : Организация;
        …
relation Сотрудник <Персона, Организация> (Должность: domen_Должности;
               Дата1: data; Дата2: data)
```

Словарь терминов

Одной из задач разрабатываемой технологии было создание гибких механизмов, позволяющих специалисту проводить тонкую настройку структуры словаря. Это выражается в возможности в значительной степени формировать структуру словарной статьи для терминов одного вида.

Типизация признаков словарной статьи. Все признаки, хранящиеся в словарной статье термина, мы условно разделили на четыре группы в зависимости от их функционального назначения.

Терминообразующие признаки служат для того, чтобы, с одной стороны, выявить термин в тексте (анализ), с другой – послужить основой для построения терминов (синтез). Для разных видов терминов набор терминообразующих признаков различен. В случае терминообразующих признаков возможность изменять структуру словарной статьи сильно зависит от вида термина (см. ниже).

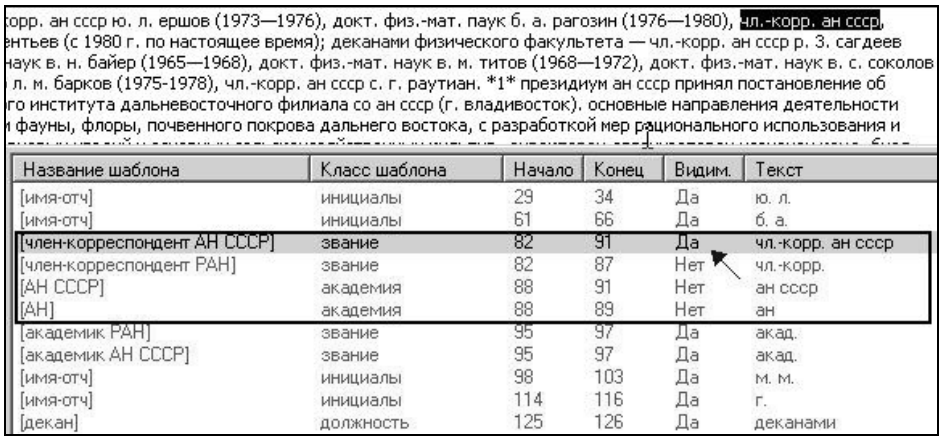
Семантические признаки приписываются терминам словаря и передаются внешним программам вместе с найденными в тексте терминами. Эта информация будет использоваться на стадии семантического анализа текста и позволит связать элементы словаря с онтологическими классами проблемной и предметной областей. Набор признаков и их тип определяется пользователем при создании и наполнении словаря.

Семантическая информация может быть выражена в словаре по-разному. Во-первых, это семантический класс (или несколько классов), к которому приписывается термин словаря. Во-вторых, это атрибуты, добавляемые пользователем в словарную статью терминов. Наличие и тип значения атрибута фиксируются в структуре словарной статьи для всех или выделенной группы терминов (класса) словаря. В-третьих, это связи между элементами словаря – такие, как отношения синонимии, омонимии, часть-целое и пр.

Статистические признаки накапливают статистическую информацию о появлении термина в обрабатываемых текстах. Часто такие признаки служат для наполнения словарей. Так, при создании ИнС, как правило, изначально имеется большая выборка ресурсов, размеченная и соотнесенная с тематическими разделами, и, используя классические методы обучения, можно сразу получить начальное наполнение словаря, которое в противном случае пришлось бы вводить вручную многочисленным специалистам. Помимо этого, наличие в словаре такой информации позволяет использовать статистические методы классификации для определения общей тематики документа.

Если в случае семантических признаков пользователь имеет полную свободу в формировании словарной статьи, то в случае статистических терминов такая свобода отсутствует. Пользователь может только задавать иерархию рубрик или тем, согласованных с обучающей выборкой, в соответствии с которой будет автоматически накапливаться статистика, и при желании вручную устанавливать веса терминов в той или иной теме, которые будут при дальнейшей обработке иметь приоритетное значение по сравнению с автоматически получаемыми значениями.

Еще одна важная группа признаков, упрощающая работу со словарями внешним программам, – это динамические признаки, которые появляются у терминов после того, как они найдены в тексте документа в результате словарного анализа текста. К ним относятся такие признаки, как *статус обновления* статистических параметров термина, *видимость* найденных терминов после словарной обработки текста (рис. 1), *позиция* найденного термина в тексте и др.


орр. ан ссср ю. л. ершов (1973—1976), докт. физ.-мат. наук б. а. рагозин (1976—1980), чл.-корр. ан ссср,
ентьев (с 1980 г. по настоящее время); деканами физического факультета — чл.-корр. ан ссср р. з. сагдеев
наук в. н. байер (1965—1968), докт. физ.-мат. наук в. м. титов (1968—1972), докт. физ.-мат. наук в. с. соколов
л. м. барков (1975-1978), чл.-корр. ан ссср с. г. раутиан. *1* президиум ан ссср принял постановление об
го института дальневосточного филиала со ан ссср (г. владивосток). основные направления деятельности
фауны, флоры, почвенного покрова дальнего востока, с разработкой мер рационального использования и

Название шаблона	Класс шаблона	Начало	Конец	Видим.	Текст
[имя-отч]	инициалы	29	34	Да	ю. л.
[имя-отч]	инициалы	61	66	Да	б. а.
[член-корреспондент АН СССР]	звание	82	91	Да	чл.-корр. ан ссср
[член-корреспондент РАН]	звание	82	87	Нет	чл.-корр.
[АН СССР]	академия	88	91	Нет	ан ссср
[АН]	академия	88	89	Нет	ан
[академик РАН]	звание	95	97	Да	акад.
[академик АН СССР]	звание	95	97	Да	акад.
[имя-отч]	инициалы	98	103	Да	м. м.
[имя-отч]	инициалы	114	116	Да	г.
[декан]	должность	125	126	Да	деканами

Рис. 1. Пример разбора цепочки вложенных терминов

Найденные термины представляются в виде объектов, у которых в данном случае имеются динамические признаки *Начало*, *Конец*, *Видимость* и *Текст*. Границы терминов *АН*, *АН СССР*, *член-корреспондент РАН* (элемент *РАН* необязательный) включены в интервал, определяемый границами термина *член-корреспондент АН СССР*, поэтому для этих терминов признак видимости отрицателен. Отрицательное значение признака *видимость* означает, что термин не будет принимать участие в дальнейшем анализе. К этой же группе можно отнести и те признаки, которые влияют на значения динамических признаков. Например, *значимость* термина позволяет управлять параметром видимости. Этот атрибут имеет следующие значения:

- самостоятельно незначимый термин всегда «невидим» и используется только в составе других терминов (например, аббревиатуры в составе других названий – *НИИ*, *СО*);
- самостоятельно значимый термин имеет положительную видимость, если не входит в состав другого термина, и отрицательную видимость в противном случае (это означает, что строка текста, покрываемую термином, строго вложена в часть текста, покрываемую другим термином; к таким терминам относится большинство предметных терминов);
- универсально значимый термин является самостоятельно значимым, но не влияет на видимость вложенных в него терминов (например, разрывные термины или термины, характеризующие целый текстовый сегмент);
- абсолютно значимый термин всегда имеет положительную видимость независимо от того, входит он в состав другого термина или нет (например, такие жанровые термины, как обращение, заключительная реплика и т. п.).

Типизация терминов. Нами были выделены несколько видов терминов, для каждого из которых реализован отдельный словарный компонент [Сидорова, 2005].

Лексема, или однословный термин, представляет собой слово во всей совокупности его форм и значений. Словарная статья лексемы содержит следующие терминообразующие признаки:

- нормальная форма;
- основа – неизменяемая при склонении или спряжении часть лексемы (у некоторых лексем основа может быть пустой, тогда лексема определяется парадигмой);
- парадигма – набор псевдофлексий, или изменяемых частей слова;
- морфологический класс (класс, в частности, включает часть речи и словообразующие морфологические признаки лексемы);
- тип – слово универсального словаря, предсказание, слово служебного словаря и т. п.

Набор морфологических признаков, включаемых в словарную статью лексемы, может легко настраиваться пользователем, являющимся специалистом в данной области.

Термин-словокомплекс – устойчивое терминологическое сочетание, характерное для выбранной предметной области. Наиболее распространенными структурами здесь являются сочетания существительного с прилагательным, существительного с существительным в косвенном падеже, существительного с другим существительным в качестве приложения. Имеются также многословные термины, иногда состоящие из трех и более слов. Структура словарной статьи словокомплекса фиксирована и не может быть расширена пользователем. Словарная статья словокомплекса (рис. 2) содержит следующие терминообразующие признаки: нормальная форма, список составляющих терминов, правило, согласно которому образуется данное сочетание.

Лексическая конструкция позволяет описывать несловарные единицы, имеющие регулярную структуру, например номер телефона, дата, инициалы, сокращения, а также специфические термины предметной области, отсутствующие в универсальном словаре русского языка или имеющие сложную структуру. Лексическая конструкция в словаре – это множество фрагментов текста произвольной сложности (в общем случае, разрывных), представляющее собой список альтернатив, связываемых с определенной строковой конструкцией (используется механизм вложенных описаний, опциональность, условия и т. д.).

Организация словарной информации. Для задачи анализа сообщений хроник было разработано два словаря: словарь предметной лексики, включающий словарь фамилий ученых, словарь имен и словарь общей лексики, связанной с деятельностью организаций; словарь лексических конструкций, включающий перечень организаций, списки званий и должностей, сокращения и другие служебные конструкции.

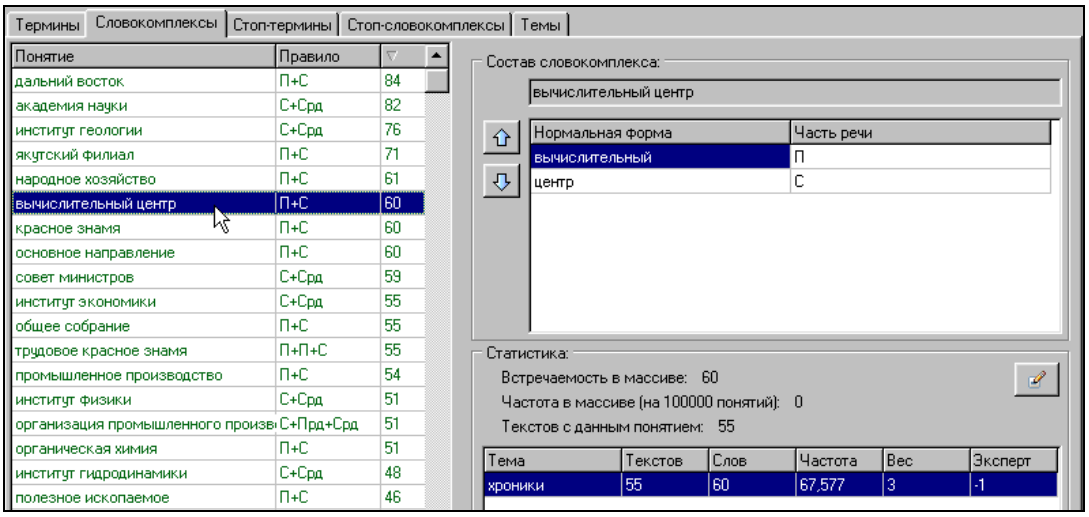


Рис. 2. Фрагмент словаря словокомплексов

С целью облегчения создания словарей было осуществлено автоматическое начальное наполнение. Для первого словаря применялись классические методы обучения, использующие универсальный морфологический словарь. Для второго, был разработан модуль, который по набору опорных терминов (*институт*, *филиал*, *президиум* и т. п.) и списку аббревиатур (*АН*, *РАН*, *СО РАН* и т. п.) формировал шаблоны вида

```
[ИСИ СО РАН] =
      институт… систем информатики( [ершова]) [СО РАН]
      иси [СО РАН]
[ПиНИИ СО РАН] =
      проект… и научно-исследовательск… институт… [СО РАН]
```

Для согласования с опорным термином левой части наименований (для определения неизменяемой части слов) использовался морфологический словарь. В дальнейшем эксперт вручную исправлял ошибки в аббревиатурах, устанавливал эквивалентность названий, отмечал необязательные фрагменты названий, формировал иерархию шаблонов и т. п.

Здесь полезно упомянуть некоторые моменты, связанные с организацией доступа к словарной информации. В предлагаемой технологии применяется xml-представление словарных данных, которое позволяет осуществлять следующее:

- согласование данных (можно указать семантическую эквивалентность словарных классов или атрибутов из разных словарей);
- расширение содержательного наполнения словарей (можно объединить элементы разных словарей в одну синонимичную группу, в дальнейшем планируется устанавливать произвольные связи между терминами);
- более точно специфицировать используемые в словаре признаки для внешней программы или внутреннего обработчика, если необходимая возможность не поддерживается словарным компонентом (например, в одном из компонентов система семантических классов и рубрик для классификации задаются идентичным образом, и для того чтобы отличить рубрику от класса, во внешнем файле просто перечисляются все индексы рубрик).

Таким образом, используемое частичное xml-представление необходимо только для согласования словарей и использования результатов словарного анализа, что позволяет достаточно эффективно применять данный механизм.

Описание фактов

Рассматривая задачу согласования предметного словаря с онтологией ИнС, можно констатировать, что элемент онтологии редко однозначно описывается с помощью одного термина, устойчивого словосочетания или другой словарной единицы. Описание обычно «размазано» по тексту, причем часто по нескольким предложениям.

Одним из способов решения этой проблемы является поиск и формирование в процессе анализа фактов из словарных единиц (или других фактов) в соответствии с заранее заданной схемой, которая явным образом связана с элементом онтологии и фиксирует правила формирования ИО по найденному факту. С технологической точки зрения необходимо иметь два компонента:

- конструктор схем фактов, который предоставит пользователю доступ к словарям и онтологии и обеспечит корректность схем;
- модуль сборки фактов, который по сформированным пользователем схемам, будет осуществлять анализ текста.

Для организации взаимодействия между этими независимыми компонентами был разработан промежуточный xml-формат описания схем фактов. Рассмотрим пример простой схемы, представленной в промежуточном формате:

```
<Scheme Name="" ArgCount="2" Segment="Предложение">
  <Argument ObjectType="TERMIN" ClassName="фио" />
       <Condition AttrName="фио-тип" Data="фам" Operation="=" />
  <Argument ObjectType="TERMIN_ALEX" ClassName="инициалы"/>
  <ConditionStruct Contact="CONTACT_ABSOLUTE"/>
  <Result Type="CREATE" ObjectType="IOBJECT" ClassName="Персона" >
       <Rule AttrName="Инициалы" Resource="Arg2" FromAttrName="Value" />
       <Rule AttrName="Фамилия"  Resource="Arg1" FromAttrName="Name" />
  </Result>
</Scheme>
```

Данная схема имеет два аргумента, которые описывают элементы из словарей разного типа (словарь предметной лексики и словарь Алекс). Структурное ограничение определяет контактность терминов в тексте. В результате будут формироваться объекты класса *Персона* с двумя определенными атрибутами:

```
<Scheme Name="Отношение_Сотрудник" ArgCount="2" Segment="Предложение">
  <Argument ObjectType="IOBJECT" ClassName="Организация"
            TypeCompare="EQUAL"/>
  <Argument ObjectType="IOBJECT" ClassName="Персона"
            TypeCompare="EQUAL"/>
  <ConditionStruct Contact="CONTACT_OBJECT"/>
  <Result Type="CREATE" ObjectType="IOBJECT_REL" ClassName="Сотрудник">
      <Rule AttrName="НазвОрг" Resource="Arg1"
            FromAttrName="Аббревиатура"/>
      <Rule AttrName="ИмяПерсоны" Resource="Arg2"
            FromAttrName="ПолноеИмя"/>
  </Result>
</Scheme>
```

Данная схема позволяет сформировать отношение *Сотрудник*. Результатом применения этих схем к фрагменту текста «*Ленинская премия присуждена группе ученых, среди которых кандидаты физ.-мат. наук Н. Н. Ефимов и Д. Д. Красильников (Институт космофизических исследований и аэрономии Якутского филиала СО АН СССР)*» стало:

***11**. СЛОЙ: позиция 193, объектов 3*
ТЕРМИН-ШАБЛОН: [имя-отч] <инициалы>,
Правая граница: 198, Значение: д. д.
ОБЪЕКТ: ***Персона*** <красильников, [имя-отч]>
Инициалы = д. д.
Фамилия = красильников
ПолноеИмя = красильников д.д.
ОБЪЕКТ: ***Сотрудник*** <Организация, Персона>
НазвОрг = [ИКФИА СО АН СССР]
ИмяПерсоны = Красильников

***12**. СЛОЙ: позиция 199, объектов 1*
ТЕРМИН-СЛОВО: красильников (фио)
мр,им,ед,фам,од
***13**. СЛОЙ: позиция 202, объектов 2*
ТЕРМИН-ШАБЛОН: [ИКФИА СО АН СССР]
<институт>, Правая граница: 223,
ОБЪЕКТ: ***Организация*** <[ИКФИА СО АН СССР]>
Имя = Институт космофизических исследований и аэрономии им. Ю.Г.Шафера СО АН СССР Аббревиатура = [ИКФИА СО АН СССР]

Модель документа

Каждый документ в зависимости от его типа или жанра имеет определенную структуру текста, которая формально представляется с помощью иерархии сегментов. Сегмент – это фрагмент текста, удовлетворяющий определенным условиям. Формально сегменты характеризуются либо «полиграфическими» элементами (абзац, строка и т. п.), либо определенной лексикой и имеют структурную организацию (состав и позиция относительно других фрагментов), а также могут реализовываться в рамках формальных сегментов других типов.

В архиве хроник сообщение имеет простую структуру, поэтому используются только «полиграфические» сегменты: предложение, абзац, скобочная структура, перечисление. При построении этих сегментов учитываются следующие два условия: сегмент может пересекаться с сегментом того же типа только границами; выбирается минимальный из возможных сегментов.

Архитектура подсистемы обработки документов

Технология, обеспечивающая обработку текста, содержит компоненты позволяющие, с одной стороны, формировать базу знаний специалистам, с другой стороны, обеспечивать автоматическое применение полученных от специалистов знаний в процессе обработки документов (рис. 3).

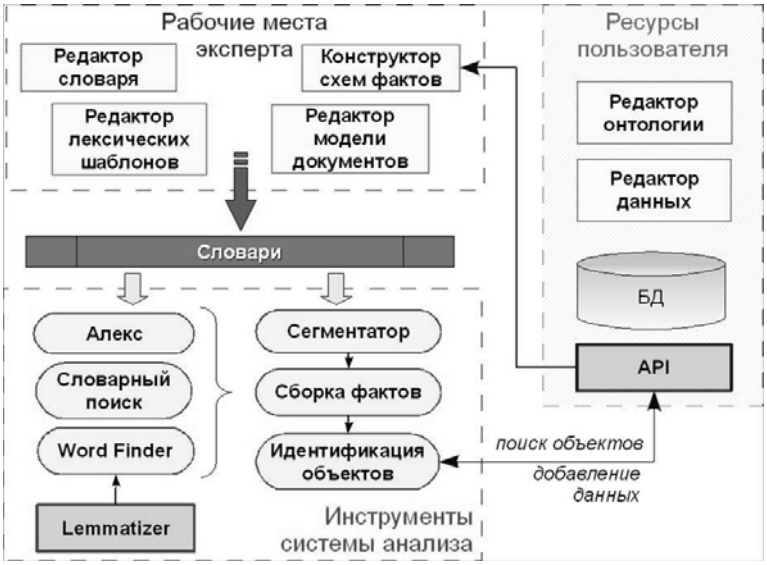


Рис. 3. Архитектура подсистемы анализа документов

Разработанная технология включает четыре функциональных компонента:

- словарный компонент, включающий АРМы настройки словарей разных типов, и исполняемые модули, реализующие основные методы словарной обработки текста;
- сегментатор, включающий АРМ настройки модели документа на основе формальных сегментов, и модуль поиска сегментов по их формальным описаниям;
- конструктор схем фактов (АРМ) и исполняемое ядро, обеспечивающее анализ текста документа начиная со словарного анализа (с помощью словарного компонента), сегментации и заканчивая процессом сборки фактов по схемам;
- модуль формирования контента, т. е. семантической сети ИО, отражающих содержание документа.

Подсистема работает в двух режимах: режим настройки базы знаний и режим обработки документа. В режиме настройки эксперт с помощью АРМов формирует базу знаний, ориентированную на использование в определенной ИнС. При создании словарей для того, чтобы автоматически извлекать ключевую лексику заданной предметной области, может быть использована подборка документов.

В режиме обработки на вход модулю анализа поступает текст документа, который передается лексическому процессору словарного компонента. Словарный компонент осуществ-

ляет сегментацию и извлечение ключевых терминов и передает результаты модулю сборки фактов. Модуль сборки фактов осуществляет поиск фактов в тексте на основании подгружаемой очереди схем фактов. Результатом работы модуля сборки фактов является множество информационных объектов. Модуль формирования контента документов идентифицирует и уточняет параметры полученных объектов, сравнивая их с информационными объектами, хранящимися в БД ИнС, формирует информационный объект документа и его контент и передает результаты в БД.

Качество работы подсистемы анализа оценивается как расхождение автоматически сгенерированного контента документа с контентом, построенным экспертом. Оценка осуществляется на массиве документов. Для оценки качества анализа применяются принятые в области автоматической обработки текстов показатели полноты и точности [Хорошевский, 2006].

Заключение

Создание подобной технологии всегда связано с поиском компромисса между настраиваемостью (декларативностью) и эффективностью анализа текста. Мы предоставили возможность пользователю практически полностью формировать лингвистическую базу знаний, однако многие вопросы, связанные с решением узкоспециальных лингвистических задач – таких, как снятие омонимии, разрешение анафоры и т. п., решаются на программном уровне.

Для оценки качества работы сервисов, создаваемых с помощью предложенной технологии, требуется дальнейшая практическая апробация.

Список литературы

Боровикова О. И., Загорулько Ю. А. Организация порталов знаний на основе онтологий // Тр. междунар. семинара Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии». Протвино, 2002. Т. 2. С. 76–82.

Васильев И. А., Тузовский А. Ф. Структура системы управления знаниями // Тр. междунар. симпозиума «Информационные и системные технологии в индустрии, образовании и науке». Караганда: Изд-во КарГТУ, 2003. С. 286–288.

Ершов А. П. К методологии построения диалоговых систем: феномен деловой прозы // Ершов А. П. Избр. тр. Новосибирск: ВО «Наука», 1994. С. 314–330.

Рубашкин В. Ш. Семантический компонент в системах понимания текста // Тр. Десятой национальной конф. по искусственному интеллекту с международным участием КИИ-2006. М: Физматлит, 2006. Т. 2. С. 455–463.

Сидорова Е .А. Технология разработки тематических словарей на основе сочетания лингвистических и статистических методов // Тр. междунар. конф. Диалог'2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. С. 443–449.

Хорошевский В. Ф. Оценка систем извлечения информации из текстов на естественном языке: кто виноват, что делать // Тр. Десятой национальной конф. по искусственному интеллекту с международным участием КИИ-2006. М: Физматлит, 2006. Т. 2. С. 464–478.

Материал поступил в редколлегию 26.08.2008

E. A. Sidirova

Factographic Text Analysis Tools in Information Systems Based on Ontologies

The technology intended for development of services of factographic analysis of text resources is considered. The knowledge base includes four components: ontology of subject domain, dictionary, formal text structure of documents and schemes of facts. Each scheme describes the structure of facts and rules to connect terms and ontology elements. The tools for creation of knowledge base and automatic usage of expert knowledge for text processing are determined.

Keywords: natural language processing, linguistic ontology, information extraction, text analysis technology.