СЕМАНТИЧЕСКИЙ ПОДХОД К АНАЛИЗУ ДОКУМЕНТОВ НА ОСНОВЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ[1]

SEMANTIC APPROACH TO ANALYSIS OF DOCUMENTS BASED ON ONTOLOGY OF SUBJECT DOMAIN

Е.А. Сидорова *(lena@iis.nsk.su)*

Ю.А. Загорулько

И.С. Кононенко

Российский НИИ искусственного интеллекта
ИСИ СО РАН им. А.П. Ершова, Новосибирск

Целью рассматриваемого подхода является извлечение из текста фактов, которые связывают найденные в тексте словарные лексические объекты и сопоставляют их понятиям онтологии. Анализ основан на предварительно создаваемых экспертом схемах, описывающих структуру фактов и учитывающих как семантическую, так и синтаксическую сочетаемость элементов каждого факта.

Введение

Задача разработки информационных систем, таких как интеллектуальные системы документооборота или информационные порталы знаний, является одной из самых актуальных на сегодняшний день. Часто она рассматривается в контексте создания хранилищ документов и их систематизации с целью облегчения поиска необходимой информации. Несмотря на важность этих вопросов, возможностей, предоставляемых существующими информационными системами, оказывается недостаточно для интеллектуальной организации деятельности: во-первых, в постоянно разрастающемся архиве становится трудно (практически невозможно) найти нужную информацию; во-вторых, данные часто дублируются и противоречат друг другу.

Современные информационные системы должны быть способны решать весь комплекс задач, связанных с управлением потоком входящих «сырых данных» – автоматическую классификацию и автоматическое индексирование текстов, оперативное и адекватное распределение новой информации среди пользователей, передачу и хранение данных в электронном архиве и последующий поиск в нем по содержанию.

Для решения этих задач разрабатывается технология автоматического анализа текста деловых или научных документов в информационных системах с ограниченной предметной областью (ПО). Эта технология должна обеспечить корректное добавление новых данных (документов) в информационное пространство системы и поддерживать содержательный поиск на основе онтологий. Она позволит осуществлять настройку базы знаний информационной системы как в момент ее создания, так и в ходе ее эксплуатации [1].

Представление знаний и данных

В технологии анализа текстовых документов выделяется три компонента знаний:

- онтология, включающая в себя *понятия и отношения ПО*; с точки зрения анализа онтология описывает данные, которые необходимо извлечь из текста и поместить в базу данных системы;
- предметный словарь (тезаурус), содержащий *термины*, с помощью которых в тексте могут представляться понятия и отношения онтологии;
- информационное наполнение системы или база данных.

Данные в системе представлены как множество разнотипных информационных объектов (ИО), которые представляют собой описание объектов предметной области и в совокупности образуют информационное наполнение системы. Каждый ИО определяется некоторым элементом онтологии (понятием или отношением) и, являясь экземпляром данного элемента, имеет заданную экспертом структуру с фиксированным набором атрибутов.

Любой ИО может быть рассмотрен в трех разных аспектах – структура, контент и контекст. Структура объекта может характеризоваться как набором собственных атрибутов и связей, так и описанием формальной структуры его содержания. Контент описывает информационное содержание объекта с помощью понятий и отношений, заданных в онтологии ПО, и представляет собой набор информационных объектов.

[1] Работа выполняется при финансовой поддержке РФФИ (проект № 04-01-00884)

Контекст в отличие от контента рассматривает ИО как единое целое и не зависит (явно) от его содержания. Контекст характеризует окружение объекта и определяется набором связей с другими объектами. Например, контекст может формироваться на основе следующих отношений:

Часть – отношение, отражающее связь ИО с охватывающим ИО (например, статьи со сборником статей);
Автор – отношение, связывающее документ с персоной, ее написавшей;
Издатель – отношение, связывающее книгу с организацией-издателем;
Информационный ресурс – отношение, по которому можно получить URL документа.

Технология анализа подразумевает работу с теми ИО, содержание которых определяется текстом. Такие ИО мы будем называть *текстовыми ресурсами*. Для того, чтобы представить в информационной системе все три описанные выше аспекта текстового ресурса, требуется:

- описать понятия (классы), которым соответствуют текстовые ресурсы;
- определить формальную структуру содержания для каждого класса текстовых ресурсов;
- задать схемы фактов, задающие правила извлечения содержательных объектов из текста.

Подход к описанию текстовых ресурсов мы рассмотрим на примере информационного объекта Документ.

Документ как информационный объект

В предлагаемом подходе анализируемые документы являются информационными объектами и описываются в онтологии некоторым понятием, например, понятием *Документ*. Текст, представляющий содержание объектов класса Документ (или любого другого класса, описывающего текстовый ресурс), анализируется с целью извлечения значимой информации или контента.

Контент документа представляет собой набор информационных объектов и их связей, описание которых встретилось в тексте документа. Для того, чтобы связать контент с документом вводится специальное отношение, позволяющее указывать для каждого экземпляра отношения (в том числе и атрибутного) индекс документа, в тексте которого он найден. (При этом найденные ИО привязывать не требуется, т.к. идентификатором ИО в тексте всегда выступает наличие хотя бы одного атрибутного отношения, связывающего сам объект с его наименованием.)

При анализе документа используется формальное представление структуры его текста, которая зависит от типа или жанра документа.

В соответствии с [2] текст в электронной форме имеет по крайней мере три уровня формальной структуры – физический, логический и жанровый. Первый представляет презентацию текста на странице, например, с помощью тегов или таблицы стилей. Ко второму уровню относятся такие элементы как текст, абзац, строка, предложение и т.п. Третий уровень представлен разбиением текста на жанровые части, например, текст делового письма [3] имеет следующие жанровые разделы: заголовок (отправитель, адресат, резюме и обращение), основной раздел (текст письма, примечания и приложения) и подпись.

Любую формальную структуру текста будем называть *сегментом* и описывать с помощью маркеров. Маркер задается списком альтернативных элементов m, где элементом m_i может быть:

1) любой символ или строка;
2) лексический объект, полученный после лексического анализа, задаваемый
   - либо классом (семантический класс, грамматический класс, служебный класс);
   - либо конкретным идентификатором или названием (для слова – это нормальная форма; для лексической конструкции, описываемой шаблоном – это название шаблона);
3) сегмент другого типа.

Построение сегмента осуществляется на основании следующих ограничений:

- single – сегмент не пересекается с сегмента ми того же типа, частный случай этого ограничения – отсутствие вложенности;
- min – выбирается минимальный из возможных сегментов на данном участке;
- max – выбирается максимальный из возможных сегментов на данном участке.

Схема факта

Иерархии классов понятий и заданные на них семантические отношения позволяют представить структуру высказывания из предметной области в виде *факта*, множество которых составляет пропозициональное содержание документа.

Информационное содержание системы представлено объектами или экземплярами понятий онтологии, поэтому в данном подходе задачей анализа является извлечение только тех фактов, которые позволяют выявить такие объекты (и их свойства и отношения) из текста. Т.е. можно рассматривать факт, как средство представления контента документа в ИС. Декларативное описание структуры факта и условий его выявления будем называть *схемой факта*.

Схема факта задает множество аргументов факта (для нашей системы будем пока рассматривать только унарные и бинарные факты), где аргументом может быть:

- понятие онтологии;

- объект или класс Тезауруса;
- факт (тип факта);
- ИО документа, текст которого анализируется.

Для того, чтобы применить схему факта ее аргументы должны удовлетворять заданным ограничениям на сочетаемость аргументов. Выделяются семантические и структурные ограничения (см. пп.0,0).

С точки зрения результата выделяются *динамические* и *статические* схемы. В результате применения динамической схемы Факта создается новый объект (ИО или факт), появление которого может послужить основанием для применения другой схемы. Результатом применения статической схемы является изменение уже существующего объекта, выступающего в качестве одного из аргументов, или ИО документа. В общем случае результатом является множество объектов или отношений найденных на заданном участке текста.

F1: Объект Исследования (памятник) +Место(географическое место) => создание
Объект-найден-в(памятник, геогр. место)
F2: Работа(работа) + Объект(объект строительства) => создание
Функция(работа, объет строительства, ВидДеятельности: строительство)
F3: Отправитель(Организация) +Функция.ВидДеятельности =>редактирование
Документ(ВидДеятельности: Функция.ВидДеятельности)

Семантические ограничения

Семантическое ограничение – это ограничение на семантический класс аргументов факта. Ограничение явно задает пару сочетаемых элементов, где элемент либо класс (наследуемый от класса аргумента), либо словарный термин, либо факт (если аргумент – факт).

Для каждой схемы Факта, описанной в онтологии, может быть сгенерирована таблица семантических сочетаний, которую должен заполнять эксперт. Назначение этой таблицы заключается в следующем:

- сужение множества вариантов сочетаний текстовых единиц;
- учет влияния аргументов друг на друга (например, уточнение семантического класса);
- уточнение атрибутов результирующего объекта.

Работа(класс) + СтрОбъект(класс) => УтРабота: Строительство
"разработка"(термин) + ПрирРесурс(класс) => УтРабота: Природопользование
"разработка"(термин) + Документ(класс) => УтРабота: Документопроизводство

Структурные ограничения

Помимо семантических ограничений, необходимо учитывать ограничения других языковых уровней, такие как синтаксические и жанровые ограничения.

Для каждой схемы факта мы будем задавать дополнительные условия на ее аргументы:

- условие на сегмент, т.е. в рамках сегмента какого типа должны располагаться аргументы;
- взаиморасположение аргументов в тексте (контактность, пре- и постпозиция, приоритетность позиции при многовариантности выбора);
- наличие синтаксических условий (валентности терминов, предложно-падежные сочетания и т.п.);
- правила образования сочетания (однородность, проективность, максимальная связность).

Проверка синтаксической сочетаемости может осуществляться либо сопоставлением грамматических признаков терминов, либо построением локального синтаксического дерева зависимостей [4].

Рассмотрим пример задания структурных ограничений:

Факт *(a1:Работа, a2:Объект):*
- условие на сегмент: Предложение;
- проверять валентность терминов класса Работа;
- проверять синтаксическую сочетаемость;
- искать однородные термины;
- соблюдать правило проективности;
- приоритетно положение Объекта справа от Работы.

«Для завершения монтажа<1> оборудования <2> и системы автоматики<3> с учетом её доработки<4>, проведения заводских испытаний <5> и подготовки к отгрузке<6> 2-й дизельной электростанции<7> заводу потребуется около 2-х месяцев»

Полученные факты:
1. <1> [монтаж] - <2> [оборудование]
2. <1> [монтаж] - <3> [сау]
3. <4> [доработка] - <4> [оборудование]
4. <4> [доработка] - <5> [сау]
5. <5> [заводские испытания] - <2> [оборудование]
6. <5> [заводские испытания] - <3> [сау]
7. <5> [заводские испытания] - <7>[электростанция]
8. <6> [отгрузка] - <7> [электростанция]

Общая схема анализа

Архитектура системы (см. Рис.1), реализующей описанный подход, включает четыре основных компонента: ядро, словарную подсистему, редакторы онтологии, схем фактов и формальных структур текста, подсистему взаимодействия с БД.

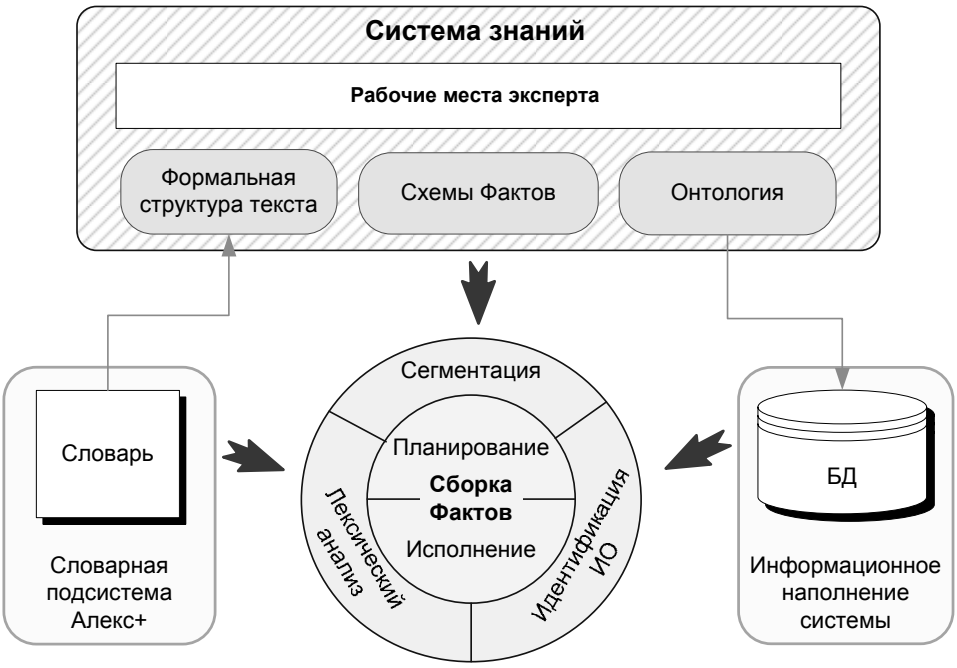

Рис. 1. Архитектура системы анализа текста на основе схем Фактов

Ядро системы обеспечивает сборку фактов по описаниям, созданным с помощью редактора. Словарная подсистема [5] обеспечивает создание словаря и предварительный этап обработки текста (сегментацию, лексический и морфологический анализ). В качестве редактора онтологии и модуля взаимодействия с БД используется компонент, реализованный в рамках проекта по созданию порталов знаний [6].

Сегментация текста

Существуют два вида сегментации текста – первичная и жанровая.

В процессе первичной сегментации осуществляется разбиение линейного представления текста на строковые объекты, оформленные как сегменты и упорядоченные в соответствии с порядком их встречаемости в тексте.

Жанровая сегментация осуществляется после лексического анализа на основе лексических объектов, маркирующих тот или иной жанровый сегмент.

Механизм сегментации реализуется с помощью системы Алекс [2], входящей в качестве подсистемы в словарный компонент предлагаемой технологии.

Лексический анализ

Лексический анализ осуществляет извлечение словарных объектов из набора упорядоченных строковых объектов, полученного после первичной сегментации текста. Здесь под словарным объектом понимаются либо лексическая конструкция, описанная с помощью системы Алекс, либо слово или словокомплекс, заданные в словаре.

В задачи данного этапа входит:

- применение лексических шаблонов;

- осуществление морфологического анализа и сборки словокомплексов;
- выделение жанровых сегментов.

Результатом работы является упорядоченный список объектов со следующим набором параметров: название (нормальная форма слова или словокомплекса, имя шаблона), позиция в тексте, значение (главное слово в синсете, извлеченное числовое значение и т.п.), грамматический класс и набор значений словоизменительных морфологических признаков для слов, набор семантических классов, статистические характеристики.

Сборка фактов

Механизм сборки фактов включает два этапа: планирование и исполнение. Причем если этап исполнения повторяется для каждого документа, то планирование осуществляется предварительно на основании заданных экспертом схем фактов.

Планирование

Задачами планирования являются:

a) Генерация исполняемых правил на основе схем фактов. Такие правила мы будем называть – *исполнителями*. Исполнители включают набор методов, которые в зависимости от типа аргументов, типа требуемого результата и набора специфических условий по-разному реализуют сборку факта заданного типа.
b) Организация очереди исполнителей. При этом необходимо учитывать три аспекта:
   - порядок создания объектов;
   - порядок и уровень вложенности сегментов (в условии на сегмент), т.е. принимать во внимание, что анализ осуществляется от самого мелкого сегмента к самому крупному по иерархии вложенности, если это не противоречит первому пункту;
   - взаимозависимость схем фактов, что требует выявления групп схем фактов, где в результате применения одной схемы, может активироваться ранее обработанная схема.

Исполнение

Во время непосредственной обработки документа, менеджер системы осуществляет последовательный вызов исполнителей из очереди. Каждому исполнителю менеджер подает на вход данные, сгруппированные по сегментам (тип сегмента задается соответствующим условием в схеме факта). Вызванный исполнитель осуществляет поиск фактов в заданном сегменте, выполняет процедурную часть (создание или редактирование объекта) и сообщает менеджеру о результате, который может послужить причиной изменения статуса других исполнителей и добавления их в текущую очередь. Процесс исполнения завершается, когда очередь становится пустой.

В зависимости от используемых алгоритмов исполнители делятся на три класса:

- исполнители, использующие таблицу семантических сочетаний;
- исполнители, использующие синтаксические правила сочетания (однородность, проективность, связность);
- все остальные.

Если задана таблица семантических сочетаний, то осуществляется предварительная оптимизация. Для этого из таблицы выбираются те сочетания, которые имеют смысл для заданного набора данных. Дальнейшая обработка рассматривает данные сочетание как отдельные схемы фактов, для которых, однако, необходимо учитывать их однородность.

При обработке однородности объекты вначале объединяются в однородные группы (группа объектов одного класса, определенного аргументом схемы факта), затем проверяется сочетаемость (семантическая и/или синтаксическая) контактных групп.

Все методы используют общий подход к разрешению омонимии на основе веса терминов и объектов. На вес оказывают влияние:

- принадлежность словокомплексу;
- сочетаемость с соседними терминами;
- вхождение в состав факта;
- статистические характеристики;
- установление референта и т.п.

Установление референта

Дальнейшая обработка заключается в уточнении полученных объектов (уточнение атрибутов) и «склеивании» одинаковых объектов, на основе использования локального контекста. Этим же целям может служить и глобальный контекст, под которым в данном подходе понимается информационное наполнение системы.

Мы выделили три задачи, решаемых с помощью глобального контекста:

- поиск в БД и идентификация объекта, найденного в тексте документа;
- разрешение текстовой омонимии (если объект уже существует в БД, то веса терминов, из которых было образовано высказывание о данном объекте, увеличиваются);

– разрешение омонимии, возникающей, когда в БД найдено несколько ИО соответствующих объекту, найденному в тексте.

Было предложено два метода разрешения омонимии.

Первый способ заключается в построении *фокусного множества ИО*, включающего все непосредственно связанные с данным ИО, и сопоставлении его с фокусным множеством объекта, найденного в тексте.

Второй способ заключается в использовании иерархии по отношению «часть-целое», в случае, когда объекты имеют сложную структуру, представленную линейными цепочками наименований, совокупность которых образует дерево (множество деревьев) ИО. Для определения такого ИО требуется восстановить иерархию вложенности объектов документа данного типа путем сравнения с эталонной иерархией ИО из БД. Каждая пара объектов, удовлетворяющая определенным требованиям порядка слов, проверяется на предмет наличия между ними отношения вложенности (с учетом транзитивности). Результирующими являются те ИО, которые соответствуют листьям полученных древесных структур.

Заключение

Предложенный подход разрабатывался на основе идеологии и методики создания комплекса ТЕОН, основные идеи которого были представлены в статье [7]. Практическая реализация основывается на методах и алгоритмах, разработанных при создании системы InDoc [1]. В ближайшее время планируется апробация предложенной технологии на задачах индексирования и классификации информационных ресурсов для порталов знаний по археологии [8] и компьютерной лингвистике.

Список литературы

1. *Загорулько Ю.А., Кононенко И.С., Сидорова Е.А., Костов Ю.В.* Подход к интеллектуализации документооборота // Информационные технологии, 2004. №11, С.2-11.
2. *Жигалов В.А., Жигалов Д.В., Жуков А.А., Кононенко И.С., Соколова Е.Г., Толдова С.Ю.*. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Труды международного семинара Диалог'2002 "Компьютерная лингвистика и интеллектуальные технологии". М.: Наука, 2002. Т.2, С.192-208.
3. *Кононенко И.С., Сидорова Е.А.* Обработка делового письма в системе документооборота // Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. М.: Наука, 2002. Т.2, С. 299–310.
4. *Гершензон Л.М., Ножов И.М., Панкратов Д.В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности. // Труды международной конференции Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии". М.: Наука, 2005. С. 97–101.
5. *Сидорова Е.А*. Технология разработки тематических словарей на основе сочетания лингвистических и статистических методов // Труды международной конференции Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии". М.: Наука, 2005. С.443-449.
6. *Zagorulko Yu., Borovikova O., Bulgakov S., Sidorova E.* Ontology-based approach to development of adjustable knowledge internet portal for support of research activity // Bull. of NCC. Ser.: Comput. Sci. 2005. Is. 23. P.45-56.
7. *Нариньяни А.С.* ТЕОН-2: от Тезауруса к Онтологии и обратно // Труды международного семинара Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии» М.: Наука, 2002. Т.1, С.199–154.
8. *Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Холюшкин Ю.П.* Концепция интеллектуального интернет-портала знаний для доступа к информационным ресурсам по археологии и этнографии // Труды VI-й международной конференции "Проблемы управления и моделирования в сложных системах". Самара: Самарский Научный Центр РАН, 2004. С. 215-220.