УДК 004.912 519.689.3

РАЗРАБОТКА ЛИНГВИСТИЧЕСКОГО ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ ОНТОЛОГИЧЕСКИХ МОДЕЛЕЙ ЗНАНИЙ

Е.А. Сидорова

Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск
Новосибирский государственный университет
E-mail: lena@iis.nsk.su

Описывается онтологическая модель знаний, ориентированная на автоматический анализ текстов в ограниченной предметной области. Рассматриваются лингвистические ресурсы и инструменты, необходимые для разработки лингвистического обеспечения информационных систем.

Ключевые слова:
Компьютерная лингвистика, онтология предметной области, лингвистическая онтология, тезаурус, лингвистические ресурсы, анализ текста.

Key words:
Computational linguistics, ontology of subject domain, linguistic ontology, thesaurus, linguistic resources, text analysis.

Введение

Разработка новых методов и моделей представления знаний является основным направлением развития современных информационных систем (ИС). С этой точки зрения можно выделить две типовые модели знаний.

Онтология информационной системы [1, 2] в явном виде описывает понятия и отношения предметной области (ПО) и позволяет не только повысить уровень представления и обработки данных в самой системе, но и создавать дополнительные предметно-ориентированные сервисы, в том числе и лингвистические. Модель хорошо поддерживает навигацию в соответствии с онтологией и «структурный» поиск, т. е. поиск в терминах онтологии. Однако для создания полноценных сервисов, ориентированных на автоматическую обработку текста, этих знаний недостаточно.

Наиболее распространенной моделью второго типа, поддерживающей работу с текстом, является тезаурус [3]. Тезаурус – это знания о языке в проекции на конкретную сферу деятельности, которые включают чисто лингвистическую информацию о «взаимоотношениях» лексики в данной области и, при необходимости, соотнесение с универсальной моделью языка. Такая модель хорошо поддерживает анализ и семантическую индексацию текстов, естественно-языковой поиск релевантных документов, создание аннотации и т. п. К недостаткам можно отнести то, что результаты запроса представлены текстами, а не конкретной информацией, как правило, присутствует много дублей и нерелевантных текстов. В таких системах возникает потребность в структуризации информации в соответствии с интересами пользователя.

В работе А.С. Нариньяни [4] предложена концепция ТЕОН, которая описывает системы, использующие при своем создании и функционировании органично взаимодействующие онтологию и тезаурус предметной области. Сложность организации такого взаимодействия связана с тем, что не всегда в языке есть однозначное соответствие между понятием и термином, часто понятия выражаются более сложным образом: словосочетаниями или фразами, части которых могут быть разнесены по тексту. Это связано с проблемой лексической многозначности слова [5] и требует для уточнения исследования контекста. Наличие определенных правил формирования текста не позволяет в полной мере отражать принципы преобразования текста с помощью тезауруса, требуются дополнительные лингвистические модели и ресурсы.

В работе будет рассмотрена расширенная модель представления знаний – *лингвистическая онтология*, ориентированная на разработку естественно-языковых сервисов, необходимых для создания и наполнения базы ИС, т. е. в первую очередь на автоматический анализ текстов в ограниченной предметной области. Данная модель достаточно полно отражает процесс связывания тезауруса и онтологии, при этом выделяются максимально независимые компоненты модели, что с технологической точки зрения дает все преимущества модульного подхода создания программных систем: независимая и распределенная реализация, повторное использование, прозрачность, легкость расширения и т. п.

1. Жизненный цикл базы знаний информационной системы

Разработка интеллектуальной ИС начинается с проектирования и наполнения ее базы знаний, которая в дальнейшем должна обновляться и пополняться, знания не должны терять свою актуальность. Таким образом, создание и развитие ИС с точки зрения ее знаний, информационного наполнения и ресурсов – это обязательно итеративный процесс. Мы будем рассматривать жизненный цикл онтологии ИС в контексте использования средств анализа текста для ее развития (рис. 1). Отметим, что задачи, также как и типы документов, обрабатываемые лингвистическими сервисами на разных этапах развития ИС, различны.

Рис. 1. *Общая схема разработки базы знаний информационной системы*

На начальном этапе разработки системы онтология играет важную роль при анализе требований и концептуальном моделировании. На данном этапе осуществляется проектирование базы знаний системы – формируется онтология верхнего уровня (метаонтология), фиксируются основные термины и понятия предметной области.

1. Онтологический анализ ПО начинается с создания словаря терминов, который используется при обсуждении и исследовании характеристик объектов и процессов, составляющих рассматриваемую ПО, также выделяются основные логические взаимосвязи между понятиями, которые соответствуют введенным терминам. Создание словаря можно автоматизировать, используя методы автоматического обучения на основе корпуса текстов предметной тематики. Задача извлечения предметной терминологии включает поиск как однословных, так и многословных терминов [6–8].
2. Задача автоматизации процесса извлечения экспертных знаний о ПО и ее подъязыке эффективно решается методами корпусной лингвистики, то есть путем создания и исследования специализированного корпуса текстов, представляющего собой достаточный объем снабженных экспертной интерпретацией лингвистических данных, который может выступать в роли обучающего корпуса. В состав корпуса текстов отбираются фрагменты из справочной и учебной литературы, научные статьи и рефераты, посвященные определенной тематике. Процессу семантической разметки специализированного корпуса текстов предшествует достаточно длительный предварительный этап совместной работы экспертов, лингвистов и разработчиков системы, в рамках которого происходит обмен компетенциями, выработка и согласование признаков и принципов разметки. Наличие семантически размеченного корпуса позволяет автоматизировать создание других лингвистических ресурсов, в первую очередь семантических словарей.
3. На следующем этапе рассматривается задача автоматического добавления справочной информации в базу знаний системы. Знания, хранимые в справочниках, должны иметь естественную проекцию в онтологии ПО. Справочные ресурсы – это, как правило, хорошо структурированные тексты, поэтому использование формальной жанровой модели таких ресурсов может значительно упростить процесс анализа текста и его настройку.
4. Накопление информации в ИС может осуществляться путем семантического индексирования потока документов. Онтология ПО определяет формат данных, хранимых в ИС, и, следовательно, определяет, какую именно информацию необходимо извлекать из текста документа, а какую можно проигнорировать. Семантический индекс документа представляется в виде сети объектов, являющихся экземплярами понятий и отношений онтологии. Данная семантическая сеть добавляется в базу данных ИС и преобразуется в знания, которыми в дальнейшем может оперировать система.
5. Задача поддержки актуальности индекса документов в таком представлении перестает зависеть от устаревания и изменения терминологии (от этого, конечно, зависит корректная обработка новых документов). Но задача устаревания информации в ИС остается актуальной и тесно связана с разрешением противоречий, возникающих при поступлении (в результате автоматического анализа новых документов) и идентификации новых данных, не согласующихся с информацией, уже присутствующей в системе.

Внесение изменений в онтологию предметной области (а также в онтологию верхнего уровня или метаонтологию) возможно либо при изменении требований к системе со стороны пользователя, либо при накоплении достаточного количества фактов, сигнализирующих о наличии неполноты в системе описания онтологии. Данные факты могут извлекаться из текста по специальным правилам с обязательным требованием высокой точности.

Для доступа к информации ИС разрабатываются пользовательские сервисы, такие как информационный поиск фактов или документов, содержащих определенные факты, представление кратких

рефератов просматриваемых документов, структурирование информации, полученной по поисковому запросу пользователя (рубрикация, кластеризация), и т. п. Сам запрос пользователь может оформлять либо на естественном языке в виде вопроса (вопросно-ответное взаимодействие), либо по ключевой фразе, либо заполняя определенную форму (формируя тем самым структурированный в соответствии с ПО запрос), либо используя навигационные средства, представляемые ИС.

2. Лингвистические потребности информационных систем

Таким образом, в процессе развития информационной системы возникают различные лингвистические задачи, решение которых требует разработки различных лингвистических ресурсов, а также средств их создания и применения (рис. 2).

В соответствии с потребностями ИС выделены следующие типы лингвистических ресурсов:

1) *Корпус текстов* – подборка текстов определенного жанра, тематика которых соответствует заданной ПО. Корпус может содержать лингвистическую разметку, представляющую собой информацию, полученную автоматически при анализе текстов либо приписанную экспертом вручную. Основное назначение корпуса – автоматизация создания других лингвистических ресурсов.
2) Универсальные и *предметные словари*, содержащие перечень минимальных единиц языка, терминов и устойчивых терминологических словосочетаний, используемых при описании значимой для ИС информации, а также жанровую лексику, описываемую лексическими шаблонами, для извлечения нестандартно представленной в тексте лексики. В рамках словарей определяется набор универсальных и/или специфичных для заданного подъязыка лингвистических знаний: морфологические классы, правила формирования многословных терминов и т. п.
3) Набор описаний жанровых структур текста в совокупности с логическим представлением текста образуют *модели документов*, соотнесенных с тем или иным типом текстовых ресурсов, хранящихся в ИС.
4) *Семантический словарь*, формирующий семантические признаки и отношения на лексиконе, включает целевые тезаурусы (например, справочно-информационный тезаурус, тезаурусы для анализа текста, для поддержки информационного поиска, для перевода и т. п.), а также словарь моделей управления, который ограничивает синтаксическую сочетаемость и проверяет согласованность грамматических и семантических признаков терминов (вершин синтаксических групп) в соответствии с правилами согласования и управления [9]. Эти знания могут быть заданы с разной степенью подробности в зависимости от требований и возможностей разработчиков ИС.
5) Знания о согласовании имеющихся лингвистических знаний с предметными знаниями, заданными онтологией ИС. С этой целью термины группируются в семантические группы, которые, в свою очередь, также согласуются с элементами онтологии либо непосредственно, либо в соответствии с определенной схемой (*схемой факта*) [10].
6) Результаты семантического индексирования всех текстов помещаются в *хранилище документов* и образуют единую информационную сеть объектов [11]. Хранилище должно поддерживать возможность поиска и идентификации объектов, а также придерживаться определенной стратегии для разрешения противоречий.

Задачи	Ресурсы	Средства
Извлечение терминологии	***Корпуса текстов*** ***Предметные словари***	*Средства разметки текстов* *Средства автоматического построения предметных словарей*
Обработка структурированных текстов	***Лексические шаблоны*** ***Тезаурусы*** ***Модели документов***	*Рабочие места лингвиста для формирования ресурсов*
Извлечение информации (фактов)	***Семантико-синтаксические модели*** ***Схемы фактов***	*Средства описания моделей и схем фактов* *Средства автоматического извлечения фактов*
Поддержка актуальности информации	***Семантически-индексированные хранилища документов***	*Средства идентификации фактов (данных)* *Стратегии разрешения противоречий на данных*

Рис. 2. *Лингвистические потребности информационных систем*

3. Лингвистическая модель знаний

Лингвистическая модель знаний должна включать всю совокупность лингвистических и экспертных знаний, необходимых для анализа текста на естественном языке. Особенностью предложенной модели является то, что в нее помимо онтологии и тезауруса включаются дополнительные компоненты, предназначенные для распознавания контекста терминов в тексте и последующего связывания найденных элементов с понятиями и отношениями предметной области.

Рассматривались следующие типы контекста.

- Устойчивое словосочетание (словокомплекс), характеризующееся наличием определенной синтаксической связи между контактно расположенными словами и высокой частотностью в анализируемом подъязыке.
- Фраза – к данному типу относятся неустойчивые словосочетания, для которых не задан строгий порядок слов, а также фразы, которые формируются нерегулярным образом: могут быть разрывными, используют различные элементы языковой отсылки.
- Множество связанных фраз – данный контекст характеризуется определенным линейным порядком связанных частей, использованием различных видов языковой редукции (сочинение и другие виды эллипсиса, анафора) и позволяет извлекать из текста информационные объекты со сложной структурой и связи.
- Текст – важным свойством данного контекста является жанр текста. Жанр играет важную роль в определении формальной структуры создаваемого автором текста (наличие в тексте таких разделов, как заголовок, резюме, основной текст).
- Информационное наполнение ИС, в том числе знания, извлеченные из ранее обработанных документов.

Формально лингвистическая модель знаний или лингвистическая онтология, для которой задана онтология предметной области O и ее информационное наполнение I_O, определяется пятеркой вида $<V,W,T,F,D>$, где V – словарь, включающий минимальные единицы текста – лексемы и лексические конструкции (сокращения, числа, численно-буквенные обозначения и т. п.), W – словарь устойчивых словосочетаний (словокомплексов) и наименований, T – семантический словарь (тезаурус), который устанавливает тезаурусные отношения между элементами словарей V и W (синонимия, родо-видовые связи, ассоциации и т. п.), F – множество упорядоченных наборов схем фактов, D – множество моделей документов, для каждой из которых может быть определен собственный набор схем фактов.

Технология, поддерживающая создание ИС, требует разработки целого комплекса лингвистических инструментов, направленных не только на непосредственную обработку текста, но и на разработку различных предметно-ориентированных лингвистических ресурсов.

4. Лингвистические процессоры

С технологической точки зрения можно выделить два типа лингвистических сервисов: системные сервисы, используемые для автоматического наполнения и изменения содержания контента системы, и пользовательские сервисы, предоставляющие пользователям разнообразный доступ к информации.

Полная цепочка процессов, обеспечивающая системные сервисы, включает графематический, лексический и морфологический анализ, жанровую сегментацию, поиск и извлечение фактов, идентификацию информационных объектов и формирование контента документа.

Процесс обработки текста начинается с предварительного словарного анализа, включающего сегментацию текста, морфологический, лексический и поверхностно-синтаксический анализ. В результате предварительного этапа формируется цепочка лексических объектов, а также сегментное покрытие текста, представленное сегментами разных типов (абзац, предложение, клауза, заголовок, скобки и т. п.), которые необходимо учитывать в условиях схем фактов.

В процессе основного анализа лексические объекты в соответствии со схемами фактов преобразуются в семантические объекты – гипотезы, которые в дальнейшем проверяются контекстом. При этом объект (как сущность предметной области, описывающая интересующую пользователя информацию) возникает на самом раннем этапе на основе семантической информации, представленной в словаре.

В общем виде задача семантического анализа текста, решаемого в рамках нашего подхода, может быть сформулирована следующим образом.

Для заданной пятерки $<O,F,T,\Omega,S>$, где O – онтология предметной области, F – множество схем фактов, T – текстовый фрагмент, Ω – терминологическое покрытие T, S – сегментное покрытие T, найти все семантические структуры, соответствующие онтологии O, покрывающие область T, которые можно получить в процессе применения правил из F к Ω с учетом S.

Ключевым понятием разрабатываемого подхода является схема фактов, которая фиксирует структуру языкового высказывания о факте действительности и явным образом связывает его с элементом онтологии ПО. Множество схем фактов описываются экспертами в терминах семантических категорий ПО и автоматически реализуются лингвистическим процессором.

Формально схема фактов – это тройка вида $<Arg,C_F,Res>$, где Arg – множество аргументов факта (семантические признаки словаря, понятия и отношения онтологии ПО, вспомогательные классы фактов), C_F – множество грамматических, синтаксических, семантических и структурных ограничений на сочетаемость аргументов, $Res=<s_i,P>$ – результат применения схемы факта, где s_i задает класс создаваемого или редактируемого объекта, а P – множество правил для формирования или уточнения значений атрибутов объекта.

Процесс поиска фактов можно рассматривать как с точки зрения продукционного подхода, где каждой схеме сопоставляется продукционное правило, так и с точки зрения мультиагентного подхода [11, 12], где каждому лексическому объекту, порождаемому на основе словарной информации, можно сопоставить *агента*, цель которого – исследовать контекст, представленный терминами и другими агентами, заполнить свои атрибуты и выявить возможные связи с другими агентами. В первом случае анализ текста заключается в последовательном применении правил к цепочке лексических объектов, во втором случае агенты сами ищут подходящие схемы фактов и осуществляют их проверку в параллельном режиме. И тот и другой подход имеет свои достоинства и недостатки, обсуждение которых выходит за рамки данной статьи.

Результатом работы лингвистического процессора является множество информационных объектов. В дальнейшем модуль формирования контента документов идентифицирует и уточняет параметры полученных объектов, сравнивая их с информационными объектами, хранящимися в ИС, формирует множество информационных объектов, представляющее проанализированный документ и его контент, а также устанавливает между ними необходимые связи.

Заключение

Рассмотренная в статье технология разработки лингвистического обеспечения для информационных систем поддерживает всю цепочку создания и использования лингвистических ресурсов для автоматического анализа текста и формирования контента документа. Система включает средства создания лингвистической онтологии, с помощью которых настройка процесса содержательной обработки документов может выполняться непосредственными носителями знаний – экспертами и лингвистами, не имеющими специальных навыков программирования.

Основные компоненты технологии были успешно апробированы в ряде практических приложений, служащих для поддержки научной и производственной деятельности, в частности при разработке интеллектуальной системы документооборота инвестиционной компании [13], портала знаний по археологии [14], архива «Хроники СО РАН» [15].

Работа выполнена при финансовой поддержке РФФИ (проект № 12–07–31216) и Президиума РАН (интеграционный проект СО РАН № 15/10 «Математические и методологические аспекты интеллектуальных информационных систем»).

СПИСОК ЛИТЕРАТУРЫ

1. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384 с.
2. Загорулько Ю.А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. – 2008. – Т. 44. – № 1. – С. 100–110.
3. Добров Б.В., Лукашевич Н.В. Онтологии для автоматической обработки текстов: описание понятий и лексических значений // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. «Диалог». – Бекасово, 2006. – М.: РГГУ, 2006. – С. 138–142.
4. Нариньяни А.С. ТЕОН-2: от Тезауруса к Онтологии и обратно // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. семинара Диалог'2002. – Протвино, 2002. – М.: Наука, 2002. – Т. 1. – С. 154–199.
5. Поляков В.Н. Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации // Проблемы прикладной лингвистики. Вып. 2 / отв. ред. Н.В. Васильева. – М.: Азбуковник, 2004. – С. 101–117.
6. Большаков И.А. Какие словосочетания следует хранить в словарях? // Труды Междунар. семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. – Протвино, 2002. – Т. 2. – С. 61–69.
7. Большакова Е.И., Носков А.А. Система для поиска и выделения конструкций в тексте на естественном языке // КИИ-2010: Труды XII национальной конф. по искусственному интеллекту с международным участием. – Тверь, 2010. – М.: Физматлит, 2010. – Т. 3. – С. 137–145.
8. Сидорова Е.А. Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. «Диалог». – Бекасово, 2008. – М.: РГГУ, 2008. – Вып. 7 (14). – С. 475–481.
9. Волкова И.А., Головин И.Г., Кривнова О.Ф. Компьютерный словарь моделей управления русских глаголов (экспериментальный вариант) // Диалог'98: Труды Междунар. семинара по компьютерной лингвистике и ее приложениям. – М., 1998. – С. 448–452.
10. Васильев И.А., Тузовский А.Ф. Структура системы управления знаниями // Информационные и системные технологии в индустрии, образовании и науке: Труды Междунар. симпозиума. – Караганда: КарГТУ, 2003. – С. 286–288.
11. Нариньяни А.С. ИИ и Мультиагентные технологи // Проблемы управления и моделирования в сложных системах: Труды VIII Междунар. конф. – Самара: Самарский научный центр РАН, 2006. – С. 491–497.
12. Вольман С.И., Минаков И.А., Томин М.С. Мультиагентная система интеллектуального анализа содержимого Интернет-страниц // Проблемы управления и моделирования в сложных системах: Труды VII Междунар. конф. – Самара: СНЦ РАН, 2005. – С. 403–408.
13. Загорулько Ю.А., Кононенко И.С., Костов Ю.В., Сидорова Е.А. Подход к интеллектуализации документооборота // Информационные технологии. – 2004. – № 11. – С. 2–11.
14. Сидорова Е.А., Загорулько Ю.А., Боровикова О.И. Подход к автоматизации извлечения информации из текстов по археологии // Информационные технологии в гуманитарных исследованиях. Вып. 11. – Новосибирск: Изд-во НГУ, 2006. – С. 1–8.
15. Кононенко И.С., Сидорова Е.А. Подход к извлечению фактов из текста на основе онтологии // Компьютерная лингвистика и интеллектуальные технологии: Матер. ежегодной Междунар. конф. «Диалог». – Бекасово, 2009. – М.: РГГУ, 2009. – Вып. 8 (15). – С. 451–457.

Поступила 30.10.2012 г.