УДК 004.82+004.912+81.322.2

Д. Ю. Власов [1, 2], **Д. Е. Пальчунов** [1, 2], **П. А. Степанов** [2]

[1] Институт математики СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия

[2] Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

E-mail: vlasov@academ.org;
palch@math.nsc.ru; stefan.nsk@gmail.com

АВТОМАТИЗАЦИЯ ИЗВЛЕЧЕНИЯ ОТНОШЕНИЙ МЕЖДУ ПОНЯТИЯМИ ИЗ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА [*]

Исследована проблема автоматического извлечения отношений между понятиями из текстов на естественном языке. Предложен метод извлечения отношений между понятиями при помощи лингвистических шаблонов, позволяющих гибко и компактно выделять в тексте различные лингвистические структуры.

Ключевые слова: онтология, понятие, автоматическая обработка текстов.

Введение

Целью настоящей работы является автоматизация извлечения информации, представленной в неструктурированных текстах естественного языка, для разработки онтологий различных предметных областей. Эти онтологии создаются в первую очередь для организации информационного поиска в сети Интернет.

В настоящий момент в Интернете представлен колоссальный объем информации, которая относится ко всем видам человеческой деятельности – все разделы науки и технологии, все отрасли производства и бизнеса, отдых, развлечения, культура, искусство и т. д. Можно с уверенностью сказать, что на подавляющее число вопросов ответы, которые уже известны человечеству, можно найти в документах, выложенных в Интернете.

И тем не менее до настоящего времени Интернет можно назвать скорее тайником сокровищ, чем универсальным решателем задач. Проблема состоит в том, что, даже если полный и исчерпывающий ответ на данный вопрос и содержится в Интернете, непонятно, как этот ответ извлечь.

При проведении научных исследований необходимо решать проблему получения новой научно-технической информации. Наиболее остро эта проблема стоит в тех областях, где имеется постоянный приток новой информации, и при этом информация довольно быстро устаревает, например, в области компьютерной безопасности. Наиболее перспективным источником оперативной научно-технической информации сейчас является Интернет.

На сегодняшний день имеются разнообразные специализированные поисковые системы, представляющие Интернет-ресурсы по научно-технической тематике. Это как универсальные поисковые системы – Google Scholar, Scirus, SciNet, ScienceDirect, Science Research Portal, Windows Live Academic, CiteSeer, InfoTrieve, Scientopica, HighWire Press и т. д.,

[*] Работа выполнена при поддержке Федерального агентства по образованию, грант ГК-П-1008.

так и специализированные – Zentralblatt MATH, MathSearch, EULER, ChemIndustry, Medline и пр.

Главный недостаток практически всех подобных систем заключается в том, что поисковый запрос представляет собой последовательность ключевых слов (плюс некоторые дополнительные возможности расширенного поиска). С помощью нескольких ключевых слов необходимо точно специфицировать предметную область, в которой ведется поиск.

Альтернативным методом решения задачи поиска научно-технической информации является разработка виртуальных каталогов для различных предметных областей [1; 2]. Для пользователя внешний вид виртуального каталога ничем не отличается от обычного каталога. Принципиальная разница – информация берется из всего Интернета, а не из базы данных каталога. Цель виртуального каталога – решать любую поисковую задачу пользователя, относящуюся к поиску научно-технической информации по данной области – математике, компьютерной безопасности, патентоведению и т. д. Виртуальный каталог является метапоисковой системой, которая перерабатывает запрос пользователя в формальные запросы к универсальным поисковым системам – Google, Yandex и др.

Поисковая задача, которую должен специфицировать пользователь, представляется в виде трехмерного вектора. Первая компонента – это раздел предметной области. Вторая – вид Интернет-ресурса. Третья компонента – вид поисковой задачи.

В настоящее время реализованы первые две компоненты – раздел предметной области и вид Интернет-ресурса. Спецификация раздела предметной области организована в виде каталога: имеется рубрикатор с различными уровнями вложенности – от пяти уровней и больше. Пользователь может выбрать рубрику из любого уровня вложенности, необязательно конечную. Например, он может выбрать «Алгебра и логика», «Математическая логика» или «Теория моделей». После этого поиск идет по выбранному разделу.

Рубрики структурированы отношением общее-частное (рубрика-подрубрика), но это отношение не является деревом. Одна и та же рубрика на любом уровне вложенности может быть непосредственной подрубрикой как одной рубрики, так и нескольких.

Таким образом, многообразие поисковых задач при быстром и легко понятном пользователю способе их спецификации можно достигнуть за счет разложения поисковой задачи в композицию подзадач. Мы выделяем три измерения поисковой задачи:

- предметная область, в которой ищется информация;
- вид Интернет-ресурса, в котором ищется информация;
- вид решаемой поисковой задачи.

Каждому из измерений поисковой задачи соответствует онтология, на основе которой осуществляется ее решение. Таким образом, для реализации информационного поиска по указанной схеме нам требуются три вида онтологий: онтология предметной области, онтология сети Интернет и онтология поисковых задач.

Проблемы и методы разработки онтологий

В настоящее время имеется большое количество исследований, посвященных применению онтологий для организации точного поиска информации в сети Интернет [3–7]. Эти исследования можно условно разделить на два подхода.

1. Применение онтологий для поиска информации, содержащейся в неструктурированных текстах естественного языка [1–3; 8].

2. Разработка формализмов для структурированного представления информации в Интернете таким образом, чтобы в дальнейшем был возможен ее автоматический поиск и обработка [7; 9]. Онтологии здесь необходимы для автоматической обработки структурированной информации.

Наша цель – извлекать информацию из неструктурированных текстов естественного языка. Для реализации виртуального каталога нам требуются три вида онтологий – предмет-

ной области, видов Интернет-ресурсов и поисковых задач [1; 2]. Наиболее важной из них является онтология предметной области.

Можно выделить два основных пути получения содержательного знания о предметной области, которое может быть представлено в виде онтологии этой предметной области:

- извлечение знаний из экспертов предметной области;
- извлечение онтологических знаний из текстов естественного языка, описывающих эту предметную область.

Решение о полностью ручном способе разработки онтологий имеет ряд существенных недостатков. Во-первых, такой способ требует чрезвычайно много ресурсов. Для того чтобы создать онтологии оперативно, необходима высокая концентрация ресурсов во времени. Во-вторых, для разработки качественных онтологий необходимо участие в качестве их разработчиков специалистов в предметной области высокого класса. В-третьих, в силу первых двух обстоятельств попытка полностью ручной разработки онтологий будет исключительно дорогостоящей. Как следствие, необходима автоматизация разработки онтологий. И эта автоматизация возможна. При этом речь ни в коем случае не идет о полностью автоматической разработке.

Во-первых, необходимо автоматизировать большой объем достаточно рутинной работы. Во-вторых, что наиболее важно, есть возможность использовать уже сделанный объем работ по описанию ключевых понятий, в который уже были вложены десятки человеколет работы ведущих специалистов данной предметной области. Речь идет об использовании энциклопедических словарей по различным областям знания. В разработке таких словарей, как правило, принимает участие коллектив ведущих специалистов. Работа проходит серьезную экспертизу и редактирование, в результате чего получается изложение, каноническое для специалистов в данной предметной области. Задача в этом случае состоит в извлечении требуемых онтологий из энциклопедических словарей. Она разбивается на две составляющих.

1. Нужно перевести текст словаря в структурированный вид, например, в XML-документ, снабженный необходимым набором тегов (о них подробнее – в следующих главах).

2. Из уже структурированного текста необходимо извлечь онтологическую информацию. Для решения этих задач необходима автоматизация извлечения онтологической информации из текстов естественного языка. Из энциклопедического словаря по данной предметной области можно извлечь следующую онтологическую информацию:

- набор ключевых терминов предметной области;
- отношения между ключевыми терминами;
- определения смысла ключевых терминов на естественном языке.

В частности, названия статей энциклопедического словаря являются ключевыми терминами предметной области. Из этого набора ключевых терминов можно выделить специальный поднабор терминов, которые являются названиями подобластей данной предметной области. Таким образом, мы, во-первых, формируем набор потенциальных рубрик виртуального каталога по данной предметной области и, во-вторых, порождаем множество потенциальных эвристик [1; 2] – ассоциаций к названиям рубрик.

После того, как сформирован набор ключевых терминов предметной области, необходимо установить отношения между этими терминами. Из энциклопедического словаря мы можем извлечь ряд важных онтологических отношений между ключевыми понятиями. Первое – это наиболее популярное отношение в объектно-ориентированном программировании – отношение «общее-частное». Второе – отношение ассоциации. Оно говорит, что два данных термина достаточно близко связаны друг с другом. Мы определяем, что два ключевых термина предметной области находятся в отношении ассоциации, если они входят в определение некоторого третьего термина. Отношение ассоциации является симметричным.

Очень важным является еще одно, несимметричное отношение – термин А требуется для определения термина В. Это отношение также можно извлечь из текста энциклопедического словаря: второй термин – это заголовок словарной статьи, а первый – это такой ключевой термин, который встречается в этой словарной статье. Заметим, что данное отношение

не только не является симметричным, но также не является и антисимметричным. Это означает, что возможны такие пары ключевых терминов предметной области, что первый из них используется в определении второго, а второй, в свою очередь, – в определении первого.

Отсутствие антисимметричности у указанного отношения возможно по ряду причин. Во-первых, есть разные способы определения смысла ключевых понятий. В одном из таких способов первое понятие определяется через второе, а в другом – наоборот, второе через первое. Во-вторых, может быть принципиально более сложная ситуация, когда понятия не могут быть явно определены одно через другое. Возможность такой ситуации вытекает из теоремы 5 [8] и теоремы 3 [10]. Данные теоремы говорят о существовании понятий, которые не допускают явных определений. В частности, существуют такие два понятия, что нельзя определить одно из этих двух понятий, не определяя одновременно и второе. В этом случае словарные статьи, определяющие каждое из таких понятий, должны излагать их связь с другими понятиями. Поэтому могут быть даже не два, а несколько понятий, каждое из которых обязательно должно встречаться в определении других.

Данное отношение между ключевыми понятиями предметной области – одно понятие необходимо для определения другого – с нашей точки зрения является не менее фундаментальным, чем такие широко используемые отношения, как «общее-частное» и «часть-целое». Оно является исключительно важным для основной цели разработки онтологий – представления в явном виде смысла ключевых понятий предметной области.

Принципы извлечения отношений между понятиями

Для решения задачи извлечения отношений между понятиями нужно, во-первых, выделить множество понятий заданной предметной области и, во-вторых, по возможности максимально полно и корректно найти все отношения между элементами, найденными на первом этапе. Для этого мы рассматриваем энциклопедические словари по интересующей предметной области как хранилища множества понятий и отношений между ними, содержащие эту информацию частично на уровне структурной разметки словарных статей и частично на естественном языке. В соответствии с этим можно выделить основные этапы извлечения отношений между понятиями:

1) выделение понятий как словарных статей в специализированных толковых словарях;

2) синтаксический анализ словарных статей, выделение значимых синтаксических элементов и определение отношений между понятиями как семантических связей между ссылками на понятия в словарных статьях.

Выделение понятий

Обычно словарная статья содержит как минимум две части, имеющие формально-синтаксические признаки выделения:

- *поле термина* (определяемое понятие);
- *поле определения* (тело словарной статьи, описывающее термин).

Также могут быть дополнительные поля, содержащие, например, грамматические характеристики термина, список синонимов, список ссылок на связанные словарные статьи.

Словарная статья имеет жесткую синтаксическую структуру, позволяющую автоматически разбивать текст словаря на словарные статьи. В качестве основных признаков, по которым можно разобрать тексты словарей на статьи, можно назвать: пустые строки – разделители; тип шрифта – жирный, наклонный, другой особенный; специальные символы – длинные пробелы, табуляции, дефисы, скобки и прочие специфические признаки.

В большинстве случаев структуру словарной статьи можно задать параметрическим образом, в виде шаблона, помещая в нужных местах набор признаков, отвечающих за ту или иную часть структуры словарной статьи.

Главной проблемой является наличие ошибок распознавания, приводящих к некачественному разбору словарных статей. Основные ошибки при автоматическом разборе словарных статей:

- слияние двух словарных статей в одну;
- выделение пустых (фиктивных) словарных статей;
- присоединение мусора, искажение слов в полях словарной статьи за счет некачественного распознавания.

Проблематичной задачей является также синтез информации из нескольких словарей в одну базу понятий. Необходимо разработать алгоритм, определяющий пары словарных статей, описывающих одно и то же понятие. Слияние текстов словарных статей только по совпадению их заголовков может оставить в словаре несколько статей, фактически описывающих одно и то же понятие. С другой стороны, ослабление условия полного совпадения заголовков словарных статей может привести к слиянию словарных статей, описывающих разные понятия.

Синтаксический анализ словарных статей

Автоматическое выделение содержательных отношений между понятиями невозможно без использования синтаксического анализа текста. С другой стороны, полный синтаксический разбор словарных статей является ненужным, так как порождает излишек информации, не требующийся для выделения отношений между понятиями. Поэтому разумно использовать следующий принцип: применять ограниченный синтаксический анализ для выделения отношений между понятиями, выделяющий только те (по возможности наиболее простые) лингвистические структуры в тексте, которые позволяют с достаточной степенью достоверности выделить отношения между понятиями.

В соответствии с этим принципом синтаксический анализ разбивается на несколько шагов, последовательно уточняющих синтаксическую структуру текста, содержащую информацию об отношениях между понятиями:

1) выделение лексем;
2) морфологический анализ;
3) выделение простых вспомогательных структур;
4) выделение структур согласования, управления и объектов;
5) выделение набора жестких структур, описывающих отношения между понятиями.

Выделение лексем, морфологический анализ

Для облегчения синтаксического анализа словарной статьи стандартную процедуру выделения лексем по символам пробелов, а также по набору выделенных символов-разделителей (скобки, точки, запятые и т. д.) можно дополнить выделением устойчивых комбинаций символов с сохранением семантической информации о типе этих комбинаций. Эта задача может решаться при помощи регулярных выражений, задающих типы подобных устойчивых комбинаций. Морфологический анализ лексем выполняется стандартным образом [11].

Выделение структур согласования, управления и потенциальных объектов

При определении какого-либо отношения между понятиями на естественном языке сами понятия представляют собой объекты – существительные в некоторой форме, определяемой типом отношения. Синтаксическая структура, описывающая понятие-объект, может быть сложной, представляющей собой словосочетание из нескольких слов, грамматически связанных с основным существительным при помощи связи управления, а также набора прилагательных перед главным существительным, согласованных с ним по роду, падежу и числу [12].

Следует отметить, что структура, определяющая согласование различных членов предложения в роде / числе / падеже, является важной при определении отношений между понятиями.

Выделение отношений между понятиями

Для выделения некоторого бинарного отношения между понятиями задается набор структур, представляющих собой тройки:

- первый объект отношения;
- структура, определяющая сказуемое, которое семантически задает данное отношение в естественном языке (включает в себя фиксированную основу глагола этого сказуемого);
- второй объект отношения.

При этом все три компоненты данной тройки должны быть грамматически согласованны между собой и образовывать неразрывный интервал входного текста. Достоверность выделения отношения напрямую зависит от точности задания компонент таких троек: чем точнее они определены, тем выше достоверность. С другой стороны, очевидно, что чем жестче задана лингвистическая структура в тексте, тем с меньшей вероятностью она будет реализована при определении отношения между понятиями в каждом конкретном случае.

Все операции по выделению синтаксических структур после морфологического анализа осуществляются при помощи так называемых *лингвистических шаблонов*. Лингвистические шаблоны образуют специализированный язык, предназначенный для гибкого и компактного определения синтаксических структур в тексте, с использованием грамматических и лексических характеристик слов в тексте.

Язык описания лингвистических шаблонов

Далее мы изложим язык описания лингвистических шаблонов. Для этого языка реализована программа-интерпретатор. На вход программа принимает текст на русском языке и файл с описаниями шаблонов. Также необходимо указать имя шаблона для поиска. После предварительной обработки текста – *разбиения на слова* и *морфологического анализа* – программа приступает к поиску участков текста, подходящих под указанный шаблон.

На выходе пользователь получает исходный текст в XML-формате, где найденные участки выделены специальными тегами. Пример описания шаблонов, выделяющих осмысленные структуры текста, приведен в конце данного раздела.

Грамматика языка

Язык описания шаблонов описан здесь в расширенной форме Бэкуса – Наура. В языке также предусмотрены комментарии в стиле UNIX: любая строчка, начинающаяся с символа #, считается комментарием.

Приведем базовые категории грамматики языка описания лингвистических шаблонов:

Символ ::= а|б|...|я|А|Б|...|Я|a|b|...|z|A|B|...|Z
Идентификатор ::= _ | *Символ* | [_ | *Символ*] *Идентификатор*
ПечатныйСимвол ::= любой печатный символ
Константа ::= *ПечатныйСимвол* | *ПечатныйСимвол Константа*

При описании шаблонов в качестве идентификаторов выступают имена шаблонов и изменяемых параметров, а в качестве констант – точные значения исходных форм и наиме-

нования свойств слов. На этапе морфологического анализа для каждого слова выясняются его грамматические свойства, наименования которых затем используются при поиске. Простыми примерами грамматических свойств служат грамматические значения – именительный падеж, родительный падеж, ..., первое лицо, ..., мужской род, женский род, и т. п.

Основной конструкцией языка является определение лингвистического шаблона. Любой файл описаний должен состоять из одного или нескольких определений шаблонов.

ОпределениеШаблона ::=
ИмяШаблона (*СписокАргументов*) = *ТелоОпределения*;
СписокАргументов ::=
ИмяАргумента | *ИмяАргумента*, *СписокАргументов*
ИмяШаблона ::= *Идентификатор*
ИмяАргумента ::= *Идентификатор*

Имя шаблона указывается как один из входных параметров программы – имя шаблона для поиска. Также именование шаблонов позволяет использовать их для определения новых.

Тело определения шаблона содержит выражения – условия для поиска участка текста, подходящего под этот лингвистический шаблон. Тело определения состоит из одного или нескольких выражений, записанных подряд. При поиске интерпретация такой записи производится следующим способом. Например, если тело определения некоторого шаблона состоит из двух выражений *A* и *B*, записанных подряд, то некоторый участок текста соответствует шаблону тогда и только тогда, когда этот участок представим в виде двух частей *X* и *Y*, причем *X* соответствует выражению *A*, а *Y* – выражению *B*.

Список аргументов содержит имена изменяемых параметров шаблона. Основное применение аргументов – описание лингвистических шаблонов, содержащих связи между словами. Известно, что связь согласования в словосочетаниях требует совпадения падежа, числа и рода прилагательного и существительного. В этом случае для описания шаблона «сочинительная связь» необходимо ввести три аргумента – род, число и падеж, использовать ссылки на эти аргументы в лингвистическом шаблоне, а также указать области их значений при старте программы (для аргумента «род» – грамматические значения рода, для аргумента «падеж» – падежи, для аргумента «число» – числа).

ТелоОпределения ::= *СписокВыражений*
СписокВыражений ::= *Выражение* | *Выражение СписокВыражений*

Предлагается пять типов выражений, которые можно указывать в теле определения шаблона. Рассмотрим каждый из них подробнее.

Выражение ::= *Слово* | *Шаблон* | *Повтор* | *Выбор* | *Пропуск*

Базовым является выражение для слова. Это выражение состоит из условия на основу слова и списка грамматических свойств, которыми данное слово должно обладать. Участок текста подходит под данный шаблон тогда и только тогда, когда этот участок содержит ровно одно слово, основа которого совпадает с указанной в шаблоне, а список грамматических свойств, найденный морфологическим анализатором при анализе этого слова, содержит все грамматические свойства, указанные в выражении. В качестве основы или свойства может быть указан специальный символ *, который, по аналогии с языками регу-

лярных выражений, интерпретируется как «любая основа» или «любое свойство». Также при указании основы и свойств могут быть использованы ссылки на аргументы шаблона. В этом случае гарантируется, что в рамках данного шаблона все ссылки на один аргумент будут заменены ровно одним значением, принадлежащем области определения этого аргумента.

Слово ::= < *УсловиеНаОснову* : *СписокСвойствСлова* >
УсловиеНаОснову ::= * | $*ИмяАргумента* | *Константа*
СписокСвойствСлова ::=
СвойствоСлова | *СвойствоСлова* : *СписокСвойствСлова*
СвойствоСлова ::= * | $*ИмяАргумента* | *Константа*

Второй тип выражений, участвующих в описании шаблона, основан на использовании описанного ранее шаблона. При поиске участка текста, соответствующего этому выражению, будет использован указанный шаблон. Если он имеет изменяемые аргументы, то должны быть указаны их значения – с помощью констант, ссылок на аргументы описываемого шаблона или же с использованием специального символа * в значении «любой». Также в качестве входного аргумента может быть передан список констант, что будет означать сужение области возможных значений этого входного аргумента. Гарантируется, что найденный участок текста будет соответствовать указанному шаблону с нужными значениями аргументов.

Шаблон ::= *ИмяШаблона* (*СписокПараметров*)
СписокПараметров ::= *Параметр* | *Параметр* , *СписокПараметров*
Параметр ::= * | $*ИмяАргумента* | *Константа* | *СписокКонстант*
СписокКонстант ::= *Константа* | *Константа* / *СписокКонстант*

Для описания шаблонов, содержащих несколько повторяющихся выражений, удобно использовать повторы.

Повтор ::= *Выражение* *

В этом случае программа будет искать в исходном тексте произвольное количество идущих подряд участков (возможно, ноль), соответствующих данному выражению, и интерпретировать их как один большой участок, соответствующий выражению-повтору.

При описании шаблонов могут использоваться выражения типа «Выбор», предназначенные для поиска участков текста, соответствующих одному из выражений из списка возможных.

Выбор ::= '[' *ПеречислениеВыражений* ']'
ПеречислениеВыражений ::=
Выражение | *Выражение* '|' *ПеречислениеВыражений*

Для поиска участков текста, не соответствующих ни одному выражению из некоторого списка, следует использовать выражения типа «Пропуск».

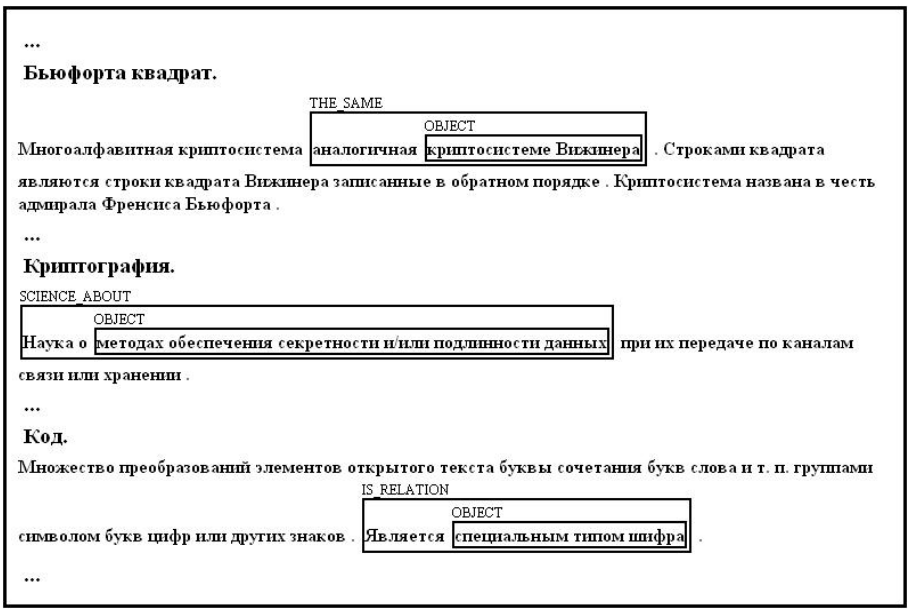
Пропуск ::= '[' *ПеречислениеВыражений*']' ~

Тогда в результате будет найдена такая область текста, что ни один ее отдельный участок не будет соответствовать ни одному выражению из указанного списка.

Выделение отношений между терминами словаря

При разработке программы был создан набор скриптов, осуществляющих выделение отношений типа общее-частное между терминами словаря. Поиск таких отношений основан на выделении предварительно заданных синтаксических структур, заданных на языке описания шаблонов, в словарных статьях.

Тестирование осуществлялось на двух словарях: математических терминов и терминов по информационной безопасности (см. рисунок). Предварительно вручную было проведено выделение отношений между терминами, а затем результаты работы были сравнены.



Анализ словаря по информационной безопасности
с помощью разработанной программы

На основе предложенного языка и шаблонов была разработана и протестирована программная система, извлекающая отношения между понятиями из текстов энциклопедических словарей. Результаты работы программы на словаре по информационной безопасности содержали 22 % ошибок, а всего было найдено 57 % всех отношений, обнаруженных человеком. Результаты работы программы на математическом словаре содержали 11 % ошибок, и было найдено 81 % всех отношений, обнаруженных человеком.

Более низкое качество работы на словаре по информационной безопасности по сравнению с математическим словарем обусловлено более неформальным стилем изложения информации. Полученная в результате тестирования оценка качества не является окончательной и может быть улучшена путем пополнения набора лингвистических шаблонов, используемых при анализе словарей.

Заключение

Два основных пути создания онтологий предметных областей – это извлечение знаний из экспертов предметной области и из текстов естественного языка, описывающих предметную область. Недостатком первого пути является достаточно высокая стоимость, необходимость затраты большого количества времени, необходимость участия большого количества специалистов – экспертов в предметной области. Полностью ручная разработка онтологий в таком случае будет весьма дорогостоящей. В работе решается задача автоматизированного извлечения онтологической информации о предметной области из текстов естественного языка.

Для этого мы используем уже сделанный объем работ по описанию ключевых понятий, в который были вложены десятки лет работы ведущих специалистов в данной предметной области. Онтологическую информацию мы извлекаем из энциклопедических словарей по различным областям знания. В разработке таких словарей принимал участие коллектив ведущих специалистов в данной области знаний, в результате чего мы имеем изложение, каноническое для специалистов в данной предметной области. Задача состоит в автоматизации извлечения онтологических знаний из энциклопедических словарей. Для этого необходимо перевести текст словаря в структурированный вид и из структурированного текста извлечь онтологическую информацию.

Для решения задачи автоматизированного построения онтологий предложен язык лингвистических шаблонов, позволяющий гибко и компактно задавать правила выделения лингвистических структур из текста на естественном языке. На этом языке описаны правила, позволяющие выделять отношения между понятиями из текста на естественном (в данном случае – русском) языке из текстов энциклопедических словарей для различных предметных областей. Для этого достаточно использовать относительно небольшое количество правил, задающих требуемые семантические отношения между понятиями.

Язык лингвистических шаблонов и набор правил для выделения конкретных отношений между понятиями (типа общее-частное) был применен к тексту математического энциклопедического словаря и энциклопедического словаря по информационной безопасности. Полученные при этом результаты оценивались по двум критериям: полноте и корректности. Полнота в обоих случаях превысила 50 %, количество ошибок не превысило 30 %, что, исходя из наших целей, можно рассматривать как вполне удовлетворительный результат.

Предложенный метод автоматизированной обработки текстов на естественном языке можно в дальнейшем развивать, уточняя и дополняя правила выделения отношений между понятиями. Можно также улучшать качество разбора, как по полноте, так и по корректности. Кроме того, можно применять язык лингвистических шаблонов и для решения других задач: для классификации предложений или выделения иных специфических лингвистических структур. Наконец, сам язык лингвистических шаблонов можно расширять и увеличивать его выразительную силу с целью его использования для более глубокого анализа текстов на естественном языке. Одной из целей дальнейшей работы является разработка универсального языка структурного анализа для обработки текстов естественного языка, который позволил бы решать целый спектр задач в этой области.

Список литературы

1. *Пальчунов Д. Е., Сидорова Е. С.* Виртуальный каталог // Тр. Всерос. конф. «Знания – Онтологии – Теории». Новосибирск, 2007. С. 166–175.

2. *Пальчунов Д. Е.* Решение задачи поиска информации на основе онтологий // Бизнес-информатика. 2008. № 1. С. 3–13.

3. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии I: Теоретико-модельная формализация онтологии и рефлексии // Философия науки. 2006. № 4 (31). С. 86–114.

4. *Fensel D.* Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag, 2001.

5. *Gómez-Pérez A., Fernandez-Lopez M., Corcho O.* Ontological Engineering with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. Springer-Verlag, 2004.

6. *Pal'chunov D. E.* GABEK for Ontology Generation // Lernen und Entwicklung in Organisationen. Learning and Development in Organizations., Beitrage zur Wissensverarbeitung. Contribution to Knowledge Organisation / Eds. Ph. Herdina, A. Oberprantacher, J. Zelger. Berlin; Wien: LIT Verlag, 2007. Bd. 2. P. 90–109.

7. Handbook on Ontologies / Eds. S. Staab, R. Studer. Berlin; Heidelberg: Springer-Verlag, 2004.

8. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. Ч. 2: Онтологии и формализация понятий // Философия науки. 2008. № 2 (37). С. 62–99.

9. *Daconta M. C., Obrst L. J., Smith K. T.* The Semantic Web. A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley Technology Publishing, 2006.

10. *Пальчунов Д. Е.* Определимость предложений языка булевых алгебр с выделенными идеалами // Вестн. Новосиб. гос. ун-та. Серия: Математика, механика, информатика. 2008. Т. 8, вып. 2. С. 62–75.

11. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение. Около 100 000 слов. М.: Рус. яз., 1997.

12. *Белошапкова В. А., Брызгунова Е. А., Земская Е. А. и др.* Современный русский язык: Учеб. для филол. спец. высших учебных заведений. 3-е изд., испр. и доп. М.: Азбуковник, 1997.

Материал поступил в редколлегию 28.05.2010

D. Yu. Vlasov, D. E. Palchunov, P. A. Stepanov

AUTOMATION OF EXTRACTION OF RELATIONS BETWEEN CONCEPTS FROM THE NATURAL LANGUAGE TEXTS

The problem of automated extraction of relations between concepts from the natural language texts is observed in the paper. We propose the method of extraction of the relations between concepts based on the specialized language of linguistic templates, which makes it possible to highlight different linguistic structures from a text in a flexible and compact way.

Keywords: ontology, concept, natural language processing.