УДК 004

ПРОЕКТИРОВАНИЕ ЛИНГВИСТИЧЕСКИХ ОНТОЛОГИЙ ДЛЯ ИНФОРМАЦИОННЫХ СИСТЕМ В ШИРОКИХ ПРЕДМЕТНЫХ ОБЛАСТЯХ

Н.В. Лукашевич[1], Б.В. Добров[2]

Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова, Москва, Россия
[1]*louk_nat@mail.ru*, [2]*dobrov_bv@mail.ru*

Аннотация

В статье представлена модель лингвистической онтологии для автоматической обработки текстов широкой предметной области, т.е. предметной области, в состав которой входят тысячи разных классов сущностей, входящих между собой в неограниченные типы отношений и ситуаций. Существенно новым в предложенной модели является набор отношений лингвистической онтологии, который специально подобран для описания широкой предметной области. Предложено использовать небольшой набор отношений, сопоставимый с набором отношений в традиционных информационно-поисковых тезаурусах. Однако были введены более строгие онтологические определения используемых отношений. Такая система отношений отражает наиболее существенные взаимосвязи между сущностями, может применяться для описания отношений между понятиями в самых разных предметных областях. Представлен пример словарной статьи из опубликованной лингвистической онтологии РуТез-lite.

Ключевые слова: *лингвистическая онтология, широкая предметная область, информационно-аналитические системы.*

Введение

В настоящее время в связи с огромными объёмами электронных документов имеется всё возрастающая потребность в обработке неструктурированной текстовой информации, повышении качества и эффективности имеющихся методов обработки текстов. В число активно развивающихся направлений обработки неструктурированной текстовой информации входят такие задачи, как собственно поиск информации, фильтрация, рубрикация и кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документа и группы документов, поиск похожих документов и дубликатов, сегментирование документов и многое другое.

Современные информационно-поисковые и информационно-аналитические системы работают с текстовой информацией в широких или неограниченных предметных областях, т.е. областях, в состав которых входят тысячи разных классов сущностей, входящих между собой в неограниченные типы отношений. Поэтому характерной чертой современных методов обработки текстовой информации в таких системах стало минимальное использование знаний о мире и о языке, опора на статистические методы учёта частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т.п. В то же время, когда подобные операции выполняет человек, ему необходимо выявить основное содержание документа, его основную тему и подтемы, и для этого обычно используется большой объём знаний о языке, о мире и об организации связного текста.

Недостаток лингвистических и онтологических знаний (знаний о мире), используемых в приложениях информационного поиска и автоматической обработки текстов, приводит к

разнообразным проблемам. Нехватка знаний приводит к нерелевантному поиску в тех случаях, если способы формулировки запросов отличаются от способов описания релевантных ситуаций в документах. Эта проблема усугубляется при обработке длинных запросов, при поиске ответов на вопросы в вопросно-ответных системах.

В последнее время всё большее значение приобретают специализированные виды информационного поиска, такие как медицинский, патентный, научный поиск, и роль знаний о предметной области в обеспечении качества работы таких информационных системах, безусловно, значительна. Кроме того, при поиске в отличных от Интернета коллекциях документов, таких, как профессиональные информационные базы, внутрикорпоративные ресурсы, отличающиеся относительно небольшим (по сравнению с Интернет) размером, несоответствие языка запроса и языка документов считается достаточно серьёзной проблемой.

В то же время внедрение в современные методы автоматической обработки текстов дополнительных объёмов знаний о языке и мире является сложной задачей. Это связано с тем, что такие знания должны описываться в специально создаваемых ресурсах (тезаурусах, онтологиях), которые должны содержать описания десятков тысяч слов и словосочетаний, иметь такие возможности, как логический вывод. При применении таких ресурсов обычно необходимо автоматически разрешать многозначность слов, т.е. выбирать их правильное значение. Кроме того, поскольку ведение любых ресурсов отстаёт от развития предметной области, необходимо развитие комбинированных методов, учитывающих как знания, так и лучшие современные статистические методы обработки текстов.

В настоящее время обсуждаются три основные парадигмы ресурсов, содержащих знания о мире и языке широких предметных областей для использования в информационно-поисковых и информационно-аналитических системах.

Самой первой широко распространенной парадигмой были традиционные информационно-поисковые тезаурусы, разработка и использование которых регламентируются национальными и международными стандартами [1, 2]. Однако такие тезаурусы создавались для ручного индексирования документов людьми-индексаторами, и в последние десятилетия, характеризующиеся резким ростом объёмов электронной информации, их роль резко снизилась.

После появления в середине 90-х годов XX века тезауруса WordNet [3], структура которого представляет собой иерархическую сеть лексикализованных понятий английского языка – синсетов, появились многочисленные работы по использованию такого рода ресурсов в качестве источника лингвистических знаний в разнообразных приложениях информационного поиска. Однако этот тезаурус создавался для проверки психолингвистической теории, и не учитывает особенностей автоматической обработки текстов, из-за чего имеется много проблем в его использовании в прикладных разработках.

Наконец, современной парадигмой компьютерных ресурсов для приложений информационного поиска являются формальные онтологии. Выдвинута концепция Семантической сети (Semantic Web) [4], базирующаяся на построении онтологических ресурсов большого объёма. Однако автоматическую обработку неструктурированных текстов на естественном языке с их неоднозначностью и неточностью трудно проводить с помощью аксиоматизированных теорий, к построению которых стремятся приверженцы формальных онтологий [5].

Поэтому для автоматической обработки текстов разрабатываются специального рода онтологии (терминологические или легкие онтологии) [6, 7], понятия в которых не определяются полностью в терминах формальных свойств и аксиом. С одной стороны, формальность описания в таких онтологиях значительно ниже, чем в формальных онтологиях. С другой стороны, формальный логический вывод на основе онтологий при анализе текста часто является необходимым, поскольку в связном тексте значительный объём информации не указы-

вается явно. Кроме того, именно в связи с применением онтологий в автоматической обработке текстов появилось понятие так называемой *лингвистической онтологии*, то есть онтологии, понятия которой в значительной мере связаны со значениями языковых единиц, терминов предметной области [8, 9].

Лингвистические онтологии охватывают большинство слов языка или предметной области и одновременно имеют онтологическую структуру, проявляющуюся в отношениях между понятиями. Поэтому лингвистические онтологии могут рассматриваться как особый вид лексической базы данных и особый тип онтологии.

Данная работа описывает модель лингвистической онтологии, предназначенной для использования в рамках автоматической обработки текстов для широких предметных областей, и конкретные ресурсы, которые разработаны на основе этой модели. В модели учитываются все три парадигмы описания знаний в широких предметных областях: информационно-поисковые тезаурусы, тезаурусы типа WordNet, онтологии. Особое внимание уделяется системе отношений между понятиями. Также описывается опубликованная лингвистическая онтология – тезаурус русского языка РуТез, который может быть полезен как источник общих знаний для создания прикладных онтологий.

1 Онтологии и лингвистические онтологии

В настоящее время наиболее распространённой формой баз знаний являются базы знаний онтологического типа. Онтологии представляют собой компьютерные ресурсы, содержащие формализованное описание фрагмента знаний о мире. Различные авторы дают разные определения для понятия онтологии [10, 11]. При всём различии к определению онтологии многие авторы соглашаются в наборе основных компонентов онтологии: классы или понятия; атрибуты (свойства); экземпляры (отдельные индивиды), отношения между классами или экземплярами; аксиомы онтологии [12].

Таким образом, формальным определением онтологий может служить следующее:

$$O = <C, E, At, R, A>,$$

где C – понятия (классы) онтологии, E – экземпляры онтологии, At – атрибуты понятий и экземпляров онтологии, R – отношения между понятиями, A – аксиомы онтологии.

Термину «онтология» удовлетворяет широкий спектр структур, представляющих знания о той или иной предметной области. В качестве в разной степени формализованных онтологий разными авторами рассматривается множество различных компьютерных ресурсов, в том числе и известных задолго до начала исследований по онтологиям таких, как рубрикаторы или тезаурусы. При этом в некоторых типах онтологий некоторые из вышеперечисленных компонентов могут быть не определены [13]. Так, структура рубрикаторов обычно не включает экземпляры и атрибуты, т.е. распространенной формальной моделью рубрикаторов является модель вида:

$$O = <C, \varnothing, \varnothing, R, A> = < C, R, A >$$

Наиболее формализованные онтологии представляют собой ***логические теории***, построенные на произвольных логических утверждениях о понятиях – аксиомах. Для описания таких формальных онтологий применяются различные логики (дескриптивные логики, модальные логики, логика предикатов первого порядка) и различные языки описания онтологий DAML+OIL, OWL, CycL, Ontolingua. Онтологии, такие как тезаурусы, рубрикаторы, понятия которых не определяются полностью в терминах формальных свойств и аксиом, иногда называются ***лёгкими онтологиями*** (lightweight ontologies) [12].

Разработчики онтологий по-разному трактуют взаимоотношения между онтологией и естественным языком. Некоторые исследователи трактуют онтологию как структуру, независимую от естественного языка, другие – как структуру, независимую от *конкретного* естественного языка, третьи вводят элементы языкового лексикона в формальное определение онтологии. Вместе с тем имеется немало подходов к построению онтологий, в которых компоненты лексикона предметной области непосредственно вводятся в формальное определение онтологии [14, 15]. Так, одной из известных формальных моделей онтологии является модель, описанная в [14]:

$$O = <L, C, F, G, H, R, A>,$$

где:

- $L = L_C \cup L_R$ – словарь онтологии, содержащий набор лексических единиц (знаков) для понятий L_C и набор знаков для отношений L_R;
- C – набор понятий онтологии;
- F и G связывают наборы лексических единиц $\{l_j\} \subset L$ с наборами понятий и отношений данной онтологии;
- H – фиксирует таксономический характер отношений (связей), при котором понятия онтологии связаны нерефлексивными, ациклическими, транзитивными отношениями $H \subset C \times C$;
- R – обозначает нетаксономические отношения между понятиями онтологии;
- A – набор аксиом онтологии.

Вместе с тем, даже в таких подходах, рассматривающих лексикон естественного языка как один из компонентов онтологической модели, ничего не говорится о методах установления соответствия между совокупностью лексических значений текстов предметной области и онтологии, лексические выражения представлены как вспомогательные элементы, называющие понятия и отношения онтологии.

Однако в установлении взаимоотношений между понятиями, словами и выражениями естественного языка имеется много проблем, начиная с того, как ввод нового понятия в онтологию связан с существующими языковыми выражениями. Кроме того, стремление к чёткой формализации отношений между понятиями в онтологии чрезвычайно трудно соблюсти в ситуации, когда необходимо создавать сверхбольшие ресурсы, и, кроме того, приводит к проблемам при установлении связей «понятие – языковое выражение».

Поэтому значительно большее распространение в приложениях автоматической обработки текстов получили вышеупомянутые «лёгкие» онтологии. Так, большое количество широко известных медицинских онтологических ресурсов представляет собой тезаурусы, не обладающие высокой степенью формализации своей структуры.

Тезаурусы представляют собой так называемые ***лингвистические онтологии***, т.е. онтологии, опирающиеся в своём построении на значения реально существующих выражений естественного языка. Наиболее известными типами тезаурусов, обсуждаемыми в качестве источников знаний для приложений обработки неструктурированной информации, являются информационно-поисковые тезаурусы и тезаурусы типа WordNet, структура которых будет рассмотрена ниже.

1.1 Информационно-поисковые тезаурусы

Информационно-поисковый тезаурус (в соответствии с определениями стандартов) – это нормативный словарь терминов на естественном языке, явно указывающий отношения меж-

ду терминами и предназначенный для описания содержания документов и поисковых запросов [1, 2].

Основными целями разработки традиционных информационно-поисковых тезаурусов являются следующие:

- обеспечение перевода естественного языка документов и пользователей на контролируемый словарь, применяемый для индексирования и поиска;
- обеспечение последовательного использования единиц индексирования;
- описание отношений между терминами;
- использование как поискового средства при поиске документов.

Основной единицей тезаурусов являются термины, которые разделяются на дескрипторы (= авторизованные термины) и недескрипторы (= аскрипторы). По своей сути дескрипторы однозначно соответствуют понятиям предметной области.

Отношения между дескрипторами обычно разделяются на два типа: иерархические и ассоциативные. Иерархические отношения обычно рассматриваются как несимметричные и транзитивные.

По ГОСТу 7.25-2001 [16] иерархические отношения обладают свойствами транзитивности и антисимметричности, которые могут быть использованы при избыточном индексировании в интересах повышения эффективности информационного поиска. Предпочтительно указывать связи между дескрипторами как отношения иерархического вида, если они обладают этими свойствами. Применяемые в информационно-поисковых тезаурусах иерархические отношения могут дифференцироваться на отдельные виды.

Основным иерархическим отношением, используемым в информационно-поисковых тезаурусах, является родовидовое отношение *выше-ниже*. Родовидовая связь устанавливается между двумя дескрипторами, если объём понятия нижестоящего дескриптора входит в объём понятия вышестоящего дескриптора. Также в качестве иерархического отношения в информационно-поисковых тезаурусах может устанавливаться отношение *часть-целое*.

Отношение ассоциации является неиерархическим. Основное назначение установления ассоциативных отношений между дескрипторами информационно-поискового тезауруса – указание на дополнительные дескрипторы, полезные при индексировании или поиске.

Основной целью разработки традиционных информационно-поисковых тезаурусов является использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования. Поэтому важно, чтобы набор дескрипторов информационно-поискового тезауруса позволял описывать тематику документов предметной области. При этом сам процесс индексирования по такому тезаурусу базируется на лингвистических, грамматических знаниях, а также знаниях о предметной области, которые имеются у профессиональных индексаторов текстов. Индексатор сначала должен прочитать текст, понять его и затем изложить содержание текста, пользуясь дескрипторами, указанными в информационно-поисковом тезаурусе. Индексатор должен хорошо понимать всю терминологию, использованную в тексте, – для описания основной темы текста ему понадобится значительно меньшее количество терминов.

Таким образом, формальную модель информационно-поискового тезауруса можно представить следующим образом:

$$ИПТ = <D_{th}, T, R_H, R_A, A_T>,$$

где:

- D_{th} – набор дескрипторов предметной области, соответствующий понятиям данной предметной области, индекс $_{th}$ означает в данном случае тот факт, что разработчики информационно-поисковых тезаурусов включают в состав дескрипторов термины предметной об-

ласти, которые необходимы для выражения основных тем документов этой предметной области;

- T – набор терминов предметной области: $D \subset T$; R_H – иерархические отношения информационно-поискового тезауруса;
- R_A – ассоциативные отношения информационно-поискового тезауруса;
- A_T – аксиомы транзитивности иерархических отношений.

Отметим, что описанная в национальных и международных стандартах модель информационно-поискового тезауруса предназначена для его использования в процессе ручного, экспертного анализа документов [1, 2]. Информационно-поисковый тезаурус, предназначенный для автоматической обработки текстов, должен содержать значительно больше информации о структуре и языке предметной области. Кроме того, отношения между терминами, указанные в тезаурусе, должны быть значительно более формализованы для использования их в автоматических режимах.

1.2 Тезаурусы типа WordNet

Лингвистический ресурс WordNet разработан в Принстонском университете США. WordNet относится к классу лексических онтологий, свободно доступен в Интернет, и на его основе были выполнены тысячи экспериментов в области информационного поиска [3]. WordNet версии 3.0 охватывает приблизительно 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч понятий, или совокупностей синонимов (synset); общее число пар лексема-значение насчитывает 200 тысяч. В разных странах предприняты усилия по созданию ресурсов для своих языков по модели WordNet.

Основным отношением в WordNet является отношение синонимии. Наборы синонимов – синсеты – основные структурные элементы WordNet. Понятие синонимии базируется на критерии, что два выражения являются синонимичными, если замена одного из них на другое в предложении не меняет значения истинности этого высказывания.

Именно определение синонимии в терминах заменимости делает необходимым разделение WordNet на отдельные подструктуры по частям речи. В состав словаря входят лексемы, относящиеся к четырём частям речи: прилагательное, существительное, глагол и наречие. Лексемы различных частей речи хранятся отдельно и описания, соответствующие каждой части речи, имеют различную структуру.

Синсет может рассматриваться как представление лексикализованного понятия (концепта) английского языка. Авторы считают, что синсет существительных представляет понятия существительных, глаголы выражают глагольные концепты, прилагательные – концепты прилагательных и т.п. Предполагается, что такое разделение соответствует психолингвистическим экспериментам, демонстрирующих, что представление информации о прилагательных, существительных, глаголах и наречиях устроено в человеческой памяти по-разному.

Большинство синсетов WordNet снабжены толкованием, подобным толкованиям в традиционных словарях, — это толкование рассматривается как одно для всех синонимов синсета. Если слово имеет несколько значений, то оно входит в несколько различных синсетов.

Каждая часть речи в WordNet имеет свой набор отношений. В различных компьютерных приложениях чаще всего используются существительные, между которыми установлены отношения синонимии, антонимии, гипонимии (гиперонимии), меронимии (*часть-целое*).

Основным отношением между синсетами существительных является родовидовое отношение, при этом видовой синсет называется гипонимом, а родовой — гиперонимом. Это транзитивное иерархическое отношение, которое может быть также названо isA-отношением. Синсет X называется гипонимом синсета Y, если носители английского языка считают нормальными предложения типа «An X is a (kind of) Y”.

Таким образом, отношения между синсетами образуют иерархическую структуру. При построении иерархических систем на базе родовидовых отношений обычно предполагается, что свойства вышестоящих понятий наследуются на нижестоящие – так называемое свойство наследования. Таким образом, существительные в WordNet организованы в виде иерархической системы с наследованием; были сделаны систематические усилия, чтобы для каждого синсета найти его родовое понятие, его гипероним.

Формальную модель ресурса типа WordNet можно представить следующим образом:

$$WN = <LC_{n,adj,v,adv}, R_{n,adj,v,adv}, S, T, M, A_n >,$$

где:

- $LC_{n,adj,v,adv} = \{LC_n, LC_{adj}, LC_v, LC_{adv}\}$ – совокупность лексикализованных понятий-синсетов, сгруппированных по разным частям речи (существительные, прилагательные, глаголы и наречия); синсет представляется собой одну лексему (слово в определённом значении) или совокупность синонимичных лексем;
- $R_{n,adj,v,adv} = \{R_n, R_{adj}, R_v, R_{adv}\}$ – наборы отношений синсетов, различающиеся для разных частей речи;
- T – текстовые выражения (слова и словосочетания), описанные в ресурсе;
- S – отношения между текстовыми выражениями и синсетами;
- M – совокупность неоднозначных текстовых выражений: $M \subset T$;
- A_n – аксиомы транзитивности и наследования, индекс n отражает тот факт, что аксиомы обсуждаются и используются в подавляющем большинстве случаев только для синсетов существительных.

В результате рассмотрения структурных особенностей информационно-поисковых тезаурусов и тезаурусов типа WordNet, можно сделать следующие выводы о сходстве и различии используемых моделей представления знаний в этих тезаурусах.

Наиболее бросающееся в глаза различие состоит в том, что информационно-поисковые тезаурусы описывают определённую предметную область, а WordNet содержит информацию о значениях общей лексики языка. Однако это различие не является принципиальным, поскольку можно строить тезаурусы типа WordNet и для конкретных предметных областей. Более значимые различия имеются в выборе единиц тезаурусов.

В информационно-поисковых тезаурусах имеется множество ограничений на включение в тезаурус языковых единиц: дескрипторы должны быть чётко отделены по смыслу друг от друга, многозначность языковых единиц практически не представлена, ограничивается глубина иерархий и т.д. Это приводит к возникновению существенного расхождения между единицами тезауруса и языковыми единицами, упоминаемыми в текстах предметной области. В тезаурусах типа Wordnet такой разницы нет: если существует слово или выражение с определёнными значениями, то оно включается в тезаурус в соответствующем количестве значений.

Если сравнивать систему отношений в стандартных информационно-поисковых тезаурусах и тезаурусах типа WordNet, то, прежде всего, нужно брать для сравнения отношения между синсетами существительных WordNet, поскольку дескрипторы информационно-поисковых тезаурусов – это обычно существительные и группы существительного. Оба типа тезаурусов имеют небольшой набор отношений, что, несомненно, объясняется разнообразием описываемых сущностей. При этом, однако, в наборе отношений информационно-поискового тезауруса имеется отношение ассоциации, которое при всей его неопределённости, позволяет лучше описать отношения между сущностями предметной области, чем отношение *часть-целое* в версии WordNet и *антонимии*.

2 Принципы разработки лингвистической онтологии для автоматической обработки текстов широкой предметной области

В предыдущем разделе было показано, что для приложений информационного поиска использовались разные лингвистические и онтологические ресурсы: информационно-поисковые тезаурусы, тезаурусы типа WordNet, формальные онтологии. Все они имеют некоторые проблемы при использовании их как ресурсов в рамках решения задач информационного поиска.

Традиционные информационно-поисковые тезаурусы создавались как инструмент для помощи человеку, их структура направлена на предоставление удобств индексатору (удаление слишком конкретных терминов, удаление близких по смыслу терминов, добавление комментариев по употреблению тех или иных дескрипторов). В связи с этим при использовании традиционных информационно-поисковых тезаурусов в автоматической обработке текстовой информации возникают существенные проблемы. В литературе предлагается использовать методы машинного обучения для проставления дескрипторов тезауруса по уже проиндексированному людьми множеству документов, создание которого представляется чрезвычайно дорогой процедурой.

Формальные онтологии, одним из провозглашаемых принципов которых является независимость от конкретного языка, сложно использовать в автоматической обработке текстов для приложений информационного поиска, поскольку для этого единицы формальной онтологии необходимо связать с единицами конкретного естественного языка. Кроме того, стремление к чёткой формализации отношений между понятиями в формальной онтологии чрезвычайно трудно соблюсти в ситуации, когда необходимо создавать сверхбольшие ресурсы, и, кроме того, приводит к проблемам при установлении связей «понятие – языковое выражение».

Ресурсы типа WordNet создаются для описания лексики языка в соответствии с лингвистическими традициями. Но любая информационная система имеет дела не только с общей лексикой, но и с конкретными предметными областями и их терминологиями. Анализируя попытки создать терминологические ресурсы на основе WordNet, следует отметить, что структура WordNet не приспособлена для описания терминологий. Раздельное описание частей речи, слишком большой набор несвязанных между собой значений, недостаточная проработанность принципов включения многословных выражений, – всё это приводит к проблемам разработки и использования терминологических ресурсов, созданных на базе модели WordNet.

Вместе с тем, в каждом из этих типов ресурсов есть те качества, которые должны присутствовать в большом лингвистическом ресурсе для информационно-поисковых приложений. Такой ресурс для автоматической обработки текстов в информационно-поисковых приложениях в широких предметных областях должен сочетать принципы различных традиций и методологий:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet (Принстонский университет);
- методологии созданий формальных онтологий.

Поскольку важно уметь описывать терминологию широких предметных областей, то необходимо использовать опыт разработки информационно-поисковых тезаурусов, а именно:

- информационно-поисковый контекст;
- единицы ресурса создаются на основе значений терминов;
- описание большого числа многословных выражений, принципы включения (невключения) многословных единиц;

- небольшой набор отношений между понятийными единицами.

Так как предполагается использовать лингвистический ресурс в автоматическом режиме обработки текстов, то необходимо использовать методологию разработки лексических ресурсов типа WordNet, в которой важны следующие положения:

- понятийные единицы создаются на основе значений реально существующих языковых выражений;
- многоступенчатое иерархическое построение лексико-терминологической системы понятий;
- принципы описания значений многозначных слов и выражений.

Из методологии разработки формальных онтологий важны следующие положения:

- разработка лингвистической онтологии как иерархической системы;
- использование для описания отношений формально определяемых отношений с формальными свойствами;
- в качестве аксиом (правил вывода) использование свойств транзитивности и наследования отношений между понятиями.

Таким образом, в результате исследований и экспериментов сформулированы принципы создания онтологических ресурсов для автоматической обработки текстов (далее ЛО – *лингвистическая онтология для автоматической обработки текстов*), которые будут изложены в следующем разделе.

3 Модель лингвистической онтологии широкой предметной области

Онтологию ЛО для широкой предметной области *D* можно формально представить следующим образом:

$$\text{ЛО}=<C, Ex, NO, R_{lo}, A_{tr,i}, S, T, M_{m,a}, L, DC>,$$

где:

- C – множество понятий онтологии, где понятие обозначает класс сущностей, обладающих одинаковыми свойствами и отношениями с другими классами сущностей;
- Ex – множество экземпляров понятий онтологии, задано отображение $E: C \rightarrow 2^{Ex}$;
- NO – множество имен понятий и экземпляров в онтологии, имена уникальны;
- R_{lo} – набор отношений между понятиями $R_{lo} \subset C \times C$, специально разработанный для автоматической обработки текстов в широких предметных областях;
- $A_{tr,i}$ – множество правил вывода, основанных на свойствах транзитивности и наследования отношений;
- T – множество текстовых входов онтологии – языковых выражений, значения которых представлены в онтологии;
- S – множество отношений между языковыми выражениями (T) и понятиями (C): $\{s\ (c_i, t_j)\}$;
- $M_{m,a}$ – множество многозначных слов и выражений из T: $M_{m,a} \subset T$; многозначные текстовые входы онтологии делятся на два подвида: M_m – текстовые входы, которые относятся к более чем одному понятию онтологии, и M_a – текстовые входы, которые многозначны, но в онтологии представлено только одно значение: $M_{m,a} = M_m \cup M_a$;
- L – множество лемматических представлений языкового выражения (т.е. представление выражения в виде последовательности слов в словарной форме, например, словосочетание *ценная бумага* представляется в лемматическом виде как *ЦЕННЫЙ БУМАГА*);
- DC – это отображение терминологического состава (TD) заданной коллекции предметной области ($Dcoll$) на текстовые входы и понятия онтологии:

$$DC: (Dcoll, TD) \rightarrow (T, C).$$

Отображение *DC* задает критерий минимальной полноты онтологии, которая должна обеспечивать покрытие терминологического состава заданной коллекции предметной области, что, собственно, и отражает суть лингвистической онтологии.

Рассмотрим компоненты модели подробнее.

Предлагаемая модель лингвистической онтологии близка к модели WordNet тем, что провозглашается необходимость подробного покрытия представленных в текстах предметной области понятий и способов их лексико-терминологического выражения в тексте. Отличие от модели WordNet заключается в том, что снижается зависимость создаваемой системы понятий от собственно языковых факторов, таких как разделение системы понятий по частям речи и синонимическая эквивалентность при замене в различных контекстах как фактор выделения понятий.

Таким образом, *лингвистическая онтология предметной области представляет собой базу знаний онтологического типа о понятийной системе и лексико-терминологическом составе предметной области.*

Единицей онтологии ЛО является понятие, как единица в системе понятий, имеющая свои специфические свойства, отличающие данную единицу от других единиц в системе понятий. Такой взгляд соответствует как современной трактовке дескрипторов в информационно-поисковых тезаурусах, так и понятий (классов) в онтологиях.

Каждое введённое понятие должно иметь однозначное имя. Именем может являться однозначное слово или словосочетание, значение которого соответствует этому понятию (т.е. один из текстовых входов понятия). Кроме того, имя может формироваться из многозначного текстового входа с сужающей значение пометой, или совокупностью синонимов, которая определяет значение однозначно. Подобная практика однозначного называния дескриптора подробно разработана в стандартах по созданию информационно-поисковых тезаурусов. Поскольку имя фиксирует особенности обозначаемого им понятия, то это соответствует и практике разработки формальных онтологий.

Каждое понятие снабжается набором текстовых входов – языковых выражений, значения которых соответствуют данному понятию. Такие языковые выражения являются между собой онтологическими синонимами. В текстах может встречаться множество вариантов текстовых входов того или иного понятия, как, например, известно о существовании множественной вариативности терминов предметной области. Эти варианты необходимо фиксировать сразу при вводе понятия, или дополнять при обнаружении в конкретном тексте, поскольку известно, что автоматические методы сопоставления терминов с учётом потенциальной вариативности приводят к снижению точности сопоставления.

В текстах предметной области значительную часть составляют слова, которые не принадлежат конкретной предметной области, могут встречаться в текстах многих предметных областях, т.е. принадлежащие общему лексикону *GL*, например, *создавать, участвовать, принимать* и многие другие. Поэтому многозначные слова, описанные в ЛО, делятся на два множества. В первое множество M_m входят выражения, которые отнесены более чем к двум понятиям в ЛО, например, *дерево как растение и дерево как материал*. Во второе множество M_a входят выражения, которые отнесены к одному понятию из ЛО, но данные слова могут иметь другое значение в *GL* (например, *стали: сталь, стать*), что отмечается специальной пометкой многозначности. Таким образом:

$$M = M_m \cup M_a,$$

$$\forall t_i ((t_i \in M_m) \rightarrow (\exists c_i, c_j \in C^D, c_i \neq c_j, (s(c_i, t_i) \wedge s(c_j, t_j)))),$$

$$\forall t_i ((t_i \in M_a) \rightarrow (\exists c_i \in C^D, c_j \in C^{GL} (s(c_i, t_i) \wedge s(c_j, t_j)))),$$

где:

- C^{GL} – система понятий общего лексикона, которые, возможно, не описаны в данной ЛО;
- C^{D} – система понятий предметной области, для которой создаётся онтология;
- s (c_i, t_i) – пара текстовое выражение и понятие онтологии, соответствующее значению этого текстового выражения.

Система отношений, используемых в лингвистической онтологии, представляет собой небольшой набор отношений, и в этом предлагаемая модель лингвистической онтологии близка к традиционным информационно-поисковым тезаурусам. Однако для установления отношений применяются строгие онтологические критерии. С каждым отношением связан свой набор аксиом, которые имеют важное значение для различных этапов автоматической обработки текстов и приложений информационного поиска.

Отношения между понятиями, описываемые в онтологическом ресурсе, предназначенном для автоматической обработки текстов в рамках информационно-поисковых приложений должны выполнять разнообразные функции.

Во-первых, эти отношения должны использоваться в классических функциях информационно-поисковых тезаурусов для расширения поискового запроса или вывода рубрики документа. Во-вторых, отношения важны для разрешения многозначности языковых единиц, включённых в ресурс, поскольку естественным методом реализации автоматической процедуры разрешения многозначности является сопоставление контекста употребления многозначной единицы в тексте и контекста соответствующего понятия в онтологическом ресурсе. В-третьих, отношения в онтологическом ресурсе могут использоваться для выявления лексической связности в текстах и использования выявленной структуры текста для улучшения качества обработки текстов.

Для реализации любой из этих функций необходимо осуществление специализированного логического вывода: встретив вхождение некоторого понятия в тексте, нужно делать многошаговые проходы по отношениям. В условиях широкой предметной области и, следовательно, необходимости создания лингвистической онтологии большой величины, для обработки текстов, не ограниченных по стилю, жанру, величине, наиболее стабильно можно опираться на те отношения, которые не исчезают, не изменяются в течение всего срока существования любого или подавляющего большинства экземпляров понятия: например, лес всегда состоит из деревьев. Поэтому в лингвистической онтологии описываются отношения только между такими понятиями c_i и c_j, которые присущи по крайней мере одному из этих понятий по определению.

В качестве аксиом используются свойства транзитивности и наследования:

$$r(c_i, c_j) \wedge r(c_j, c_k) \rightarrow r(c_i, c_k) \qquad (A_{tr})$$

$$r(c_i, c_j) \wedge r_1(c_j, c_k) \rightarrow r_1(c_j, c_k) \qquad (A_i)$$

Набор отношений ЛО, их свойства и особенности их описания будут рассмотрены в следующем разделе.

4 Модель отношений в лингвистической онтологии

Для логического вывода при обработке текстов в широкой предметной области необходимо, прежде всего, описывать наиболее существенные отношения между понятиями, сохраняющие свою значимость, надёжность в различных контекстах упоминания понятий.

Было выдвинуто предположение, которое подтвердилось в ходе экспериментов в различных предметных областях, что наиболее значимыми отношениями между понятиями являются те, которые связаны с *сосуществованием этих понятий или их экземпляров*. В резуль-

тате в модели лингвистической онтологии используются четыре основных отношения; каждое из них обладает вышеуказанными свойствами, которые будут подробно рассмотрены в следующих подразделах.

В качестве основных отношений онтологии ЛО используется следующий набор надёжных отношений: *выше-ниже, часть-целое*, отношение онтологической зависимости, обозначаемое как несимметричная ассоциация: *асц1-асц2.* Кроме того, в ограниченных случаях используется симметричная ассоциация – *асц.* Рассмотрим набор отношений более подробно.

4.1 Отношение *выше-ниже*

Родовидовое отношение *выше-ниже*, получившее своё название от традиционных тезаурусов, в нашей модели трактуется как онтологическое отношение *класс-подкласс*.

Пусть *выше*(c_i, c_j) – отношение *класс-подкласс* между понятиями c_i и c_j (c_i является видом (подклассом) c_j), $r(c_i, c_j)$ – это произвольное отношение между понятиями c_i и c_j. Тогда свойства отношения *класс-подкласс* могут быть записаны следующим образом:

1) транзитивность отношения класс-подкласс:

$$(\textit{выше}(c_i, c_j) \wedge \textit{выше}(c_j, c_k) \rightarrow \textit{выше}(c_i, c_k));$$

2) свойство наследования по отношению класс-подкласс:

$$\textit{выше}(c_i, c_j) \wedge r(c_j, c_k) \rightarrow r(c_i, c_k)$$

Исторически наиболее ранними принципами установления родовидовых отношений, используемых и в работах по искусственному интеллекту, и в компьютерной лингвистике, было использование ставших классическими диагностических высказываний. Например, если понятие c_i является видом понятия c_j, t_i является текстовым входом понятия c_i, t_j является текстовым входом понятия c_j, можно сказать, что «t_i *– это* t_j», «t_i,.. *и другие* t_j», «*к числу* t_j *относятся* t_i».

Однако позже выяснилось, что одни и те же выражения естественного языка (и, в частности, применяемые диагностические тесты) могут с онтологической точки зрения соответствовать значительно различающимся отношениям между сущностями внешнего мира, в том числе обладающими совсем другими свойствами [17]. Поэтому многие методические руководства по разработке понятийных ресурсов рекомендуют осуществлять дополнительные проверки для устанавливаемого отношения *класс-подкласс*.

Наиболее распространённой рекомендацией для проверки правильности установления отношений *класс-подкласс* является проверка принадлежности экземпляров нижестоящего понятия c_i множеству экземпляров вышестоящего понятия, что подразумевает ответ на вопрос: если объект является экземпляром одного понятия, то будет ли он обязательно (т.е. по определению) экземпляром некоторого другого понятия c_j [18], т.е.:

$$\forall c_i, c_j \in C, \forall e \in E(c_i) : \textit{выше}(c_i, c_j) \rightarrow e \in E(c_j).$$

Одной из серьезных проблем описания отношений класс-подкласс в онтологиях, и в частности в лингвистических онтологиях, применяемых в автоматической обработке текстов, является их смешение с описанием отношений «тип-роль»: от понятия-типа к понятию-роли.

Дж. Сова [19] определяет понятие-роль следующим образом: «Подтипы сущности могут быть двух видов: натуральные типы и ролевые типы, которые являются подтипами натуральных типов в конкретных образцах отношений (particular pattern of relationships). Человек, например, является натуральным типом, а учитель – это подтип человека в ситуации обучения». Сова предлагает простой тест для определения, является ли понятие ролью:

> *rt – является ролевым типом, если сущность может быть охарактеризована как rt только при рассмотрении другой сущности, действия или состояния.*

В соответствии с взглядом Дж. Совы роли ассоциируются с отношениями, но при этом они сущности, а не отношения.

В работе [17] Н. Гуарино отмечает, что тест Совы для различения типов и ролей недостаточен: например, нечто может быть охарактеризовано как автомобиль, только если оно имеет, по крайней мере, колеса и мотор, но автомобиль является типом, а не ролью. В работе [20] условие, сформулированное Совой, заменяется на условие так называемой внешней онтологической зависимости:

> *Понятие c_i называется внешне зависимым от понятия c_j, если для всех экземпляров c_i должен существовать экземпляр c_j, который не является частью или материалом экземпляра c_i.*

Авторы работы [21] замечают, что точнее это условие можно сформулировать так:

> *Экземпляр понятия c_j не должен быть внутренним для экземпляра понятия c_i, т.е. не должен быть частью, или материалом, или качеством (цвет).*

Например, понятие сын является внешне зависимым от понятия родитель, поскольку существует только в рамках семьи по отношению к своим родителям. С другой стороны, автомобиль не является внешне зависимым от какой-либо сущности, поскольку требует существования мотора, который является частью автомобиля. Таким образом, данное условие формализует определение ролей, данное Дж. Совой.

Ошибкой при описании ролей является описание их как вышестоящих понятий для типов, которые могут их занимать [17, 22]. Например, прочитав следующий фрагмент статьи Википедии о консервантах: «*Наиболее используемыми консервантами в древнем мире были поваренная соль, мёд, вино*», - лингвист, инженер по знаниям может представить эту информацию посредством отношения *класс-подкласс* между понятием «*консервант*» и понятием «*поваренная соль*». Но в широкой предметной области это приведёт к тому, что все экземпляры поваренной соли будут трактоваться как консерванты, любое упоминание поваренной соли в тексте может быть предъявлено пользователю при поиске как пример упоминания консерванта, что неверно.

Действительно, такое представление неточно описывает свойства сущностей, поскольку не каждая порция поваренной соли является консервантом, таким образом, нарушается основной принцип установления отношений *класс-подкласс*.

4.2 Отношение *часть-целое*

Отношение часть-целое является одним из самых известных и полезных в разных предметных областях. Особенностью отношения часть-целое является разнообразие его проявлений. Наиболее типичными объектами, к которым применяется это отношение, являются физические объекты. Но также это отношение может устанавливаться и между сущностями, длящимися во времени, между группами сущностей, ролями и процессами и др.

При моделировании этого отношения в компьютерных ресурсах важным вопросом является обеспечение транзитивности этого отношения. Свойство транзитивности обычно постулируется в классических аксиомах философской мереологии, однако на практике с транзитивностью возникают проблемы [23]. В связи с этим при моделировании этого отношения в конкретных компьютерных ресурсах приходится принимать специализированный набор решений. Так, например, подход к описанию отношений *часть-целое* в ресурсах типа WordNet в значительной степени отличается от традиций описаний этих отношений в информационно-поисковых тезаурусах [1, 3]. В рамках многих онтологий предлагаются свои принципы представления отношений *часть-целое* [24, 25].

При описании отношения *часть-целое* в предлагаемой модели лингвистической онтологии были сделаны усилия, чтобы обеспечить транзитивность этого отношения. Если обсуждать свойства транзитивности и наследования для отношения часть-целое в ресурсе, предназначенном для автоматической обработки текстов в информационно-поисковых приложениях, то наиболее важной операцией, которую необходимо обеспечить, является релевантность обсуждения частей обсуждению целого. То есть необходимо описывать отношения часть-целое так, что если текст или его некоторый фрагмент посвящен обсуждению части, то можно предполагать, что этот текст (или его фрагмент) будет релевантен и обсуждению целого [26].

Важным условием для обеспечения такого наследования является онтологическая зависимость существования части от существования целого. Действительно, если всё существование некоторой части связано с существованием целого, то и тексты, обсуждающие эту часть, будут иметь непосредственное отношение и к целому, даже если это целое в тексте явно не упомянуто. Зависимость части может быть двух видов:

- зависимость по существованию, когда экземпляр части нельзя отделить от экземпляра-целого (*балкон зала – зрительный зал*), т.н. неотделимые части;
- родовая зависимость, при которой существование экземпляра-части требует существование хотя бы одного экземпляра целого (*двигатель автомобиля – автомобиль*), т.е. обязательные целые [27].

Этим требованием, в частности, обеспечивается выполнение рекомендаций руководств и стандартов по разработке информационно-поисковых тезаурусов [1] в том, что описание иерархических отношений должно быть независимо от контекста их упоминания. Описание таких независимых от контекста, «надёжных» отношений в ресурсах, предназначенных для автоматической обработки текстов, имеет большое значение, поскольку в автоматическом режиме часто бывает невозможно использовать контекст для подтверждения существования того или иного отношения.

Накладывая вышеперечисленные условия установления отношения *часть-целое*, мы принимаем достаточно широкую трактовку этого отношения: между физическими объектами (*балкон зала – зрительный зал*); между регионами (*Европа – Евразия*); между веществами (*амидная группы – амиды*); между множествами (*батальон – рота*); между частями текста (*строфа – стихотворение*); между процессами (*номер представления – представление*).

Также отношения *часть-целое* устанавливаются для связей между сущностями, одна из которых является внутренней, зависимой от другой, таких как:

- характерные свойства (*водоизмещение – судно*);
- роль в процессе (*инвестор – инвестирование*) (ср. [21, 28]);
- участник сферы деятельности – сфера деятельности (*машиностроительный завод – машиностроение*).

В таком широком понимании описываемые отношения *часть-целое* в предлагаемой модели лингвистической онтологии наиболее близки к так называемым *внутренним отношениям* (*internal relations*), описанных Н. Гуарино в работе [29].

В настоящее время в ЛО используются следующие свойства отношения часть-целое:

часть $(c_1, c_2) \leftrightarrow$ *целое* (c_2, c_1);

целое $(c_1, c_2) \wedge$ *целое* $(c_2, c_3) \rightarrow$ *целое* (c_1, c_3) – транзитивность отношения;

выше $(c_1, c_2) \wedge$ *целое* $(c_2, c_3) \rightarrow$ *целое* (c_1, c_3) – наследование отношения целое по отношению *выше-ниже.*

Приведём примеры вывода на основе свойства транзитивности:

целое (*ОБВИНЯЕМЫЙ ПО ДЕЛУ, СУДЕБНОЕ ОБВИНЕНИЕ*)
$\wedge$ *целое* (*СУДЕБНОЕ ОБВИНЕНИЕ, СУДЕБНЫЙ ПРОЦЕСС)*
$\rightarrow$ *целое* (*ОБВИНЯЕМЫЙ ПО ДЕЛУ, СУДЕБНЫЙ ПРОЦЕСС);*

целое (АПТЕКА, ЛЕКАРСТВЕННОЕ ОБЕСПЕЧЕНИЕ)
$\wedge$ *целое (ЛЕКАРСТВЕННОЕ ОБЕСПЕЧЕНИЕ, МЕДИЦИНСКАЯ ПОМОЩЬ)*
$\wedge$ *целое* (*МЕДИЦИНСКАЯ ПОМОЩЬ, ЗДРАВООХРАНЕНИЕ*)
$\rightarrow$ *целое (АПТЕКА, ЗДРАВООХРАНЕНИЕ).*

В информационных системах такие цепочки часто интерпретируются следующим образом: если в тексте обсуждается *обвиняемый по делу*, то этот текст релевантен и таким темам, как *судебное обвинение, судебный процесс* и т.д.

В результате в создаваемых по данной модели лингвистических онтологиях реально работает вывод по транзитивности отношений *часть-целое*, что является новым достижением для лингвистических онтологий. В тезаурусе WordNet транзитивность отношения *часть-целое* не предполагалась, а в рекомендациях по информационно-поисковым тезаурусам это отношение сводилось к весьма узкому набору случаев, из-за чего такой вывод не мог играть значительной роли.

4.3 Отношение несимметричной ассоциации

Отношение несимметричной ассоциации asc_1-asc_2 *представляет внешнюю онтологическую зависимость* между понятиями [17, 29].

Это отношение устанавливается между понятиями c_1 и c_2, если выполняются два критерия:

- между понятиями c_1 и c_2 не могут быть установлены ни отношение *класс-подкласс*, ни отношение *часть-целое*;
- следующее утверждение является истинным: существование c_2 означает существование c_1.

Эти два условия означают, что понятие c_2 (зависимое понятие) является внешне зависимым от c_1:

$$asc_1(c_2, c_1) = asc_2(c_1, c_2).$$

Приведём примеры отношений внешней зависимости, которые представляются в модели в виде направленных ассоциаций:

- asc_2(*АВТОМОБИЛЬ, АВТОМОБИЛЬНАЯ ПРОМЫШЛЕННОСТЬ*): понятие *АВТОМОБИЛЬНАЯ ПРОМЫШЛЕННОСТЬ* зависит от понятия *АВТОМОБИЛЬ*, поскольку оно существует, если только существует понятие *АВТОМОБИЛЬ*;
- asc_2(*ДЕРЕВО, ЛЕС*): понятие *ЛЕС* существует, если только понятие *ДЕРЕВО* существует. Отметим, что в предложенной системе отношений мы не можем описать отношение между понятиями *ДЕРЕВО* и *ЛЕС* как отношение *часть-целое*, поскольку деревья могут расти во многих разных местах, не только в лесу.

Отношения онтологической зависимости применимы к различным областям, поэтому чаще всего они используются в онтологиях верхнего уровня [19, 21]. Кроме того, в работе [30] авторы обсуждают важность этого отношения в области биологии: движение клетки не может возникнуть без существования клетки. Однако впервые в описываемой модели пред-

лагается использовать отношение онтологической зависимости в лингвистической онтологии.

4.4 Группировки понятий и отношений в ЛО

Для различных приложений автоматической обработки текстов применяются некоторые группировки понятий (окрестности) и отношений (пути) в ЛО.

Для каждого понятия $c_i \in C$ может быть определена окрестность понятия $O_i \subset C$, такая, что $c_j \in O_i$, если существует набор понятий $\{c_1, \ldots c_k\}$ такой, что $r_1(c_i, c_1), \ldots r_2(c_n, c_{n+1}) \ldots r_r(c_k, c_j) \in R$, и на основе аксиом A_t, A_i выводимо отношение $r(c_i, c_j)$:

$$r_1(c_i, c_1), \ldots r_2(c_n, c_{n+1}) \ldots r_r(c_k, c_j) \mapsto r(c_i, c_j).$$

На рисунке 1 изображена схема окрестности понятия.

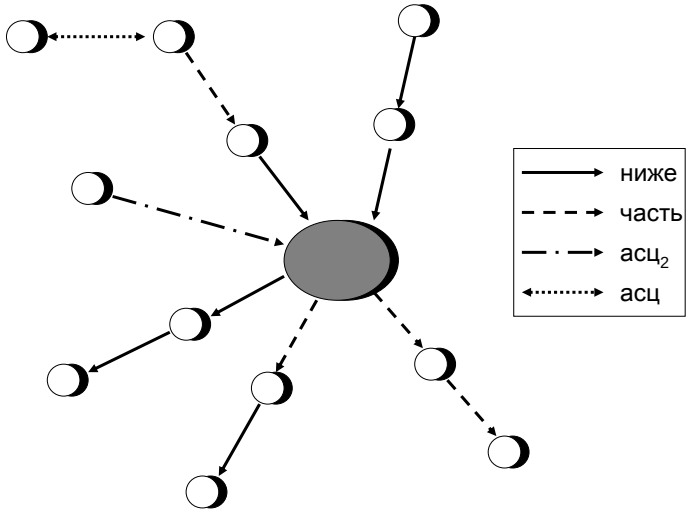

Рисунок 1 - Схема окрестности понятия лингвистической онтологии

На множестве отношений ЛО может быть введено отношение иерархии I по следующим правилам:

$$выше(c_1, c_2) \rightarrow I(c_1, c_2);$$
$$целое(c_1, c_2) \rightarrow I(c_1, c_2);$$
$$асц_1(c_1, c_2) \rightarrow I(c_1, c_2);$$
$$асц(c_1, c_2) \rightarrow I(c_1, c_2).$$

Это отношение означает, что правый элемент отношения считается более высоким по иерархии, чем левый. Для отношения симметричной ассоциации оба члена отношения равноправны.

В окрестности понятия c_i можно определить верхнюю полуокрестность O^+ и нижнюю полуокрестность O^-:

$$O^+(c_i) \cup O^-(c_i) = O(c_i);$$
$$c_j \in O^+(c_i), \text{если } c_j \in O(c_i) \wedge I(c_i, c_j);$$
$$c_j \in O^-(c_i), \text{если } c_j \in O(c_i) \wedge I(c_j, c_i).$$

Пересечение $O^+(c_i)$ и $O^-(c_i)$ может быть непустым из-за существования отношений симметричной ассоциации, входящих в обе полуокрестности. Верхняя полуокрестность понятия c_i также называется *дерево-вверх* понятия c_i, нижняя полукрестность понятия c_i – *деревом-вниз* понятия c_i.

Можно определить следующие виды путей между понятиями:

- *путь по иерархии вверх* $P_{up}(c_0, c_{00})$: от понятия c_0 к понятию c_{00} существует путь по иерархии вверх, если $c_{00} \in O^+(c_0)$;
- *путь по иерархии вниз* $P_{down}(c_0, c_{00})$: от понятия c_0 к понятию c_{00} существует путь по иерархии вниз, если $c_{00} \in O^-(c_0)$;
- *путь с перегибом вверх* $P_{updown}(c_0, c_{00})$: между понятиями c_0 и c_{00} такими, что $c_0 \notin O(c_{00})$ и $c_{00} \notin O(c_0)$, существует путь с перегибом-вверх, если существует точка перегиба – понятие c_i такое, что

$$\exists\, c_i : c_i \in O^+(c_0) \wedge c_i \in O^+(c_{00});$$

- *путь с перегибом вниз* $P_{downup}(c_0, c_{00})$: между понятиями c_0 и c_{00} такими, что $c_0 \notin O(c_{00})$ и $c_{00} \notin O(c_0)$, существует путь с перегибом-вниз, если существует точка перегиба – понятие c_j такое, что:

$$\exists\, c_j : c_j \in O^-(c_0) \wedge c_j \in O^-(c_{00}).$$

Введённые типы концептуальных путей используются в процедурах автоматического разрешения лексической неоднозначности, расширения поискового запроса, вывода рубрик по тексту [26].

Приведём примеры путей по иерархии вверх от понятия *ГОНОЧНЫЙ БОЛИД (*источник - лингвистическая онтология, тезаурус РуТез, см. раздел 5):

ГОНОЧНЫЙ БОЛИД
– ГОНОЧНЫЙ АВТОМОБИЛЬ – АВТОМОБИЛЬНЫЕ ГОНКИ - ***АВТОСПОРТ***

ГОНОЧНЫЙ БОЛИД
– ГОНОЧНЫЙ АВТОМОБИЛЬ – СПОРТИВНЫЙ АВТОМОБИЛЬ – АВТОМОБИЛЬ
– АВТОМОТОТРАНСПОРТНОЕ СРЕДСТВО – ТРАНСПОРТНОЕ СРЕДСТВО
– ***ТЕХНИЧЕСКОЕ УСТРОЙСТВО***

5 Лингвистические онтологии, созданные на основе описанной модели

Вышеописанные принципы были положены в основу разработки нескольких больших ресурсов для информационного поиска: общественно-политического тезауруса, тезауруса русского языка РуТез, онтологии по естественным наукам и технологиям ОЕНТ [31], тезауруса Банка России, Авиа-онтологии и ряда других [26].

Созданные онтологические ресурсы имеют одинаковую структуру. Они являются онтологиями, поскольку описывают понятия внешнего мира и отношения между ними. Эти ресурсы принадлежат к особому классу онтологий, так называемым лингвистическим онтологиям, поскольку введение понятий в значительной мере мотивируется значениями языковых единиц, относящихся к предметной области ресурса. В то же время они являются тезаурусами, поскольку каждое понятие связано с набором языковых выражений (слов, терминов, словосочетаний), которыми это понятие может быть выражено в тексте, – такой набор текстовых входов понятий необходим для использования онтологий для автоматической обработки текстов.

Разнообразие предметных областей, для которых созданы эти ресурсы, доказывают универсальность предложенной модели лингвистической онтологии, т.е. посредством такой модели можно описывать базовые свойства и отношения понятий, присутствующие в любой

предметной области. Объёмы созданных ресурсов демонстрируют удобство модели для быстрого наращивания ресурсов.

В настоящий момент опубликована первая версия лингвистической онтологии - тезауруса РуТез – РуТез-lite [32], созданная по предложенной модели. Опубликованная версия включает 26365 понятий, к которым приписано около 115 тысяч слов и выражений, между понятиями установлено около 108 тысяч отношений. Версия тезауруса выложена на сайте http://www.labinform.ru/ruthes/index.htm. Полная версия тезауруса РуТез включает 54 тысячи понятий, 160 тысяч слов и выражений, планируется подготовить следующие версии тезауруса к публикации.

Произведённая выкладка позволяет просматривать тезаурус по алфавиту его текстовых входов. Выбор конкретного текстового входа, например, *дерево* позволяет увидеть совокупность понятий, которым приписано данное слово, а именно к понятиям *ДРЕВЕСНОЕ РАСТЕНИЕ* и *ДРЕВЕСИНА (МАТЕРИАЛ)*. Для каждого понятия указаны полные списки текстовых входов, включающие слова разных частей речи, а также словосочетания. Так, для понятия *ДРЕВЕСНОЕ РАСТЕНИЕ* текстовыми входами являются слова и выражения: *дерево, деревце, деревцо, древесная культура, древесная порода, древесное растение, древесный, древо* (рисунок 2).

Главная О проекте Справка

А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ы Э Ю Я

ДРЕВЕСНОЕ РАСТЕНИЕ

(*ДЕРЕВО*, *ДЕРЕВЦЕ*, *ДЕРЕВЦО*, *ДРЕВЕСНАЯ КУЛЬТУРА*, *ДРЕВЕСНАЯ ПОРОДА*, *ДРЕВЕСНОЕ РАСТЕНИЕ*, *ДРЕВЕСНЫЙ*, *ДРЕВО*)

ВЫШЕ	РАСТЕНИЕ
НИЖЕ	КАКАО (ДЕРЕВО)
НИЖЕ	ЛЕСОНАСАЖДЕНИЕ
НИЖЕ	ЛИСТВЕННОЕ ДЕРЕВО
НИЖЕ	ПАЛЬМА
НИЖЕ	ПЛОДОВОЕ ДЕРЕВО
НИЖЕ	ПОДРОСТ ДЕРЕВЬЕВ
НИЖЕ	ХВОЙНОЕ ДЕРЕВО
$НИЖЕ_{В}$	ДРЕВЕСНО-КУСТАРНИКОВАЯ ПОРОДА
ЧАСТЬ	ДРЕВЕСИНА (ЧАСТЬ ДЕРЕВА)
ЧАСТЬ	ДРЕВЕСНАЯ КОРА
ЧАСТЬ	КРОНА ДЕРЕВА
ЧАСТЬ	ПОРОДА ДЕРЕВЬЕВ
ЧАСТЬ	СТВОЛ ДЕРЕВА
$АССОЦ_{2}$	ДРЕВЕСИНА (МАТЕРИАЛ)
$АССОЦ_{2}$	ЛЕСНОЙ МАССИВ
$АССОЦ_{2}$	ЛЕСОЗАГОТОВИТЕЛЬНЫЕ РАБОТЫ
$АССОЦ_{2}$	ОПАСТЬ, ОСЫПАТЬСЯ
$АССОЦ_{2}$	ПЕНЬ ДЕРЕВА

Рисунок 2 - Статья понятия *ДРЕВЕСНОЕ РАСТЕНИЕ*

Для каждого понятия указаны отношения с другими понятиями. На рисунке 2 в статье понятия *ДРЕВЕСНОЕ РАСТЕНИЕ* указаны виды деревьев, части дерева (*СТВОЛ ДЕРЕВА*,

КОРА ДЕРЕВА, *КРОНА ДЕРЕВА* и др.), а также онтологически зависимые понятия, т.е. понятия, которые не могли бы появиться, если бы в нашем мире не существовало бы деревьев: *ДРЕВЕСИНА (МАТЕРИАЛ)*, *ЛЕСНОЙ МАССИВ* (т.е. лес), *ПЕНЬ ДЕРЕВА* и др.

Тезаурус RuTez-Lite может быть получен в виде архива файлов в формате XML. Набор файлов включает в себя:

- список понятий тезауруса, для которых указана предметная область (общий лексикон, общественно-политическая область, география); также для большей части понятий имеются толкования, автоматически извлечённые из Викисловаря [32];
- список отношений между понятиями тезауруса,
- список текстовых входов тезауруса; описание текстового входа содержит лемматическое представление текстового входа, синтаксический тип (именная группа, глагольная группа и др.), главное слово именной группы [32];
- список соответствий текстовых входов понятиями тезауруса.

Заключение

В статье представлена модель лингвистической онтологии для автоматической обработки текстов широкой предметной области, учитывающая три существующие методологии создания компьютерных ресурсов. Существенно новым в предложенной модели является набор отношений лингвистической онтологии, который специально подобран для описания широкой предметной области.

Для качественного выполнения всех различных функций отношений лингвистической онтологии при автоматической обработке текстов в приложениях информационного поиска важно обеспечить многошаговый логический вывод, что может быть достигнуто на базе свойств транзитивности и наследования. Кроме того, при описании отношений необходимо добиться того, чтобы отношения были максимально «надёжными», не зависели от контекста упоминания понятия.

Для обеспечения этих свойств было предложено использовать небольшой набор отношений, сопоставимый с набором отношений в традиционных информационно-поисковых тезаурусах. Однако были введены более строгие онтологические определения используемых отношений. Такая система отношений отражает наиболее существенные взаимосвязи между сущностями, может применяться для описания отношений между понятиями в самых разных предметных областях.

Разнообразие предметных областей, для которых созданы лингвистические онтологии по предложенной модели, доказывает универсальность этой модели, её способность описывать базовые свойства и отношения понятий, присутствующие в любой предметной области. Объёмы созданных ресурсов демонстрируют удобство модели для быстрого наращивания ресурсов.

Благодарности

Данная работа осуществляется при поддержке фонда РГНФ, грант №15-04-12017.

Список источников

[1] ISO 25964-1:2011, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval / Geneva: International Organization for Standards, 2011.

[2] ANSI/NISO Z39.19-2005, Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. – Bethesda, MD: NISO Press, 2005.

[3] ***Miller, G.*** Nouns in WordNet / G. Miller // WordNet – An Electronic Lexical Database. – The MIT Press, 1998. – P. 23-47.
[4] ***Berners-Lee, T.*** The Semantic Web / T. Berners-Lee, J. Handler, O. Lassila // Scientific American - 2001. – V. 284. – No 5. – P. 28-37.
[5] ***Nirenburg, S***. What's in a symbol: Ontology, representation, and language / S. Nirenburg, Y. Wilks // Journal of Experimental and Theoretical Artificial Intelligence. - 2001. - V. 13(1). - P. 9-23.
[6] ***Ландэ, Д.В***. Подход к созданию терминологических онтологий / Д.В. Ландэ, А.А. Снарский // Онтология проектирования. - 2014. - № 2(12). - С. 83-91.
[7] ***Sowa, J.*** Building, Sharing and Merging Ontologies. - http://www.jfsowa.com/ontology/ontoshar.htm (Актуально на 07.03.2015).
[8] ***Magnini, B***. Merging Global and Specialized Linguistic Ontologies / B. Magnini, M. Speranza // Proceedings of OntoLex. - 2002. - P. 43-48.
[9] ***Veale, T.*** A context-sensitive framework for lexical ontologies / T. Veale, Y. Hao // Knowledge Engineering Review. - 2007. -Vol. 23(1). - P. 101-115.
[10] ***Guarino, N***. Ontologies and Knowledge Bases: Towards a Terminological Clarification / N. Guarino, P. Giaretta // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. - Amsterdam: IOS Press, 1995. - P. 25-32.
[11] ***Клещев, А.С***. Классификация свойств онтологий. Онтологии и их классификации / А.С. Клещев, Е.А. Шалфеева // НТИ сер. 1. - 2005. - №9. - С. 16-22.
[12] ***Corcho, O***. Roadmap to Ontology Specification Languages / O. Corcho, A. Gomez-Perez // Knowledge Engineering and Knowledge Management. Methods, Models and Tools. / Eds: R. Dieng and O. Corby. - Springer, 2000. - P. 80-96.
[13] ***Гаврилова, Т.А***. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский // СПб: Питер, 2000. - 384 с.
[14] ***Maedche, A***. Learning Ontologies for the Semantic Web / A. Maedche, S. Staab // Proceedings of Semantic Web Workshop. – Hongkong, 2001.
[15] ***Buitelaar, P***. Towards Linguistically Grounded Ontologies. The Semantic Web: Research and Applications / P. Buitelaar, Ph. Cimiano, P. Haase, M. Sintek // Proceedings of the European Semantic Web Conference. LNCS-5554. - Springer Verlag, 2009. - P. 111-125.
[16] ГОСТ 7.25.-2001. Тезаурус информационно-поисковый одноязычный: Правила разработки: структура, состав и форма представления // Система стандартов по информации, библиотечному и издательскому делу. – Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2001.
[17] ***Guarino, N***. Some Ontological Principles for Designing Upper Level Lexical Resources / N. Guarino // Proceedings of First International Conference on Language Resources and Evaluation. - Granada, Spain, 1998.
[18] ***Gomez-Perez, A***. OntoWeb. Technical Roadmap. D.1.1.2. / A. Gomez-Perez, M. Fernandez-Lopez, O. Corcho // IST project IST-2000-29243, 2001.
[19] ***Sowa, J***. Knowledge Representation: Logical, Philosophical, and Computational Foundations / J. Sowa // Brooks Cole Publishing Co., Pacific Grove, CA. 2000.
[20] ***Guarino, N.*** Evaluating ontological decisions with ONTOCLEAN / N. Guarino, C. Welty // Communications of the ACM. - 2002. - V. 45(2). - P. 61-65.
[21] ***Masolo, C.***. Social roles and their descriptions / C. Masolo**,** L. Vieu, E. Bottazzi, C. Catenacci, R. Ferrario, A. Gangemi, N. Guarino // Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning. - AAAI Press. 2004.
[22] ***Steinmann, F***. The representation of roles in object-oriented and conceptual modelling / F. Steinmann // Data and Knowledge engineering. - 2000. - V. 35. – No. 1. - P. 83-106.
[23] ***Winston, M.*** A Taxonomy of Part-Whole Relations / M. Winston, R. Chaffin, D. Hermann // Cognitive Science/ - 1989. - No. 11. - P. 417-444.
[24] ***Niles, I***. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology / I. Niles, A. Pease // Proceedings of the IEEE International Conference on Information and Knowledge Engineering.- 2003. - P. 412-416.
[25] ***Masolo, C.*** WonderWeb. Final Report / C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, L. Shneider // Deliverable D18. 2003.
[26] ***Лукашевич, Н.В.*** Тезаурусы в задачах информационного поиска / Н.В. Лукашевич. – М.: Изд-во Московского университета, 2011. – 512 с.
[27] ***Guizzardi, G.*** Ontological Foundations for Conceptual Part-Wholes Relation: The Case of Collectives and Their Parts / G. Guizzardi // Advanced Information Systems Engineering. Springer CAiSE. LNCS 6741. – Springer, 2011. - P. 138–153.

[28] ***Loebe, F***. Abstract vs. Social Roles: A Refined Top-level Ontological Analysis / F. Loebe // Proceedings of the 2005 AAAI Fall Symposium 'Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems. - AAAI Press, 2005. - P.93-100.
[29] ***Guarino, N***. The ontological Level: Revisiting 30 yers of Knowledge Representation / N. Guarino // Conceptual Modeling: Foundations and Applications. - Springer-Verlag Berlin, Heidelberg, 2009. - P. 52-67.
[30] ***Kumar, A***. The ontology of blood pressure: a case study in creating ontological partitions in biomedicine / A. Kumar, B. Smith. - http://ontology.buffalo.edu/medo/BPO.pdf (Актуально на 07.03.2015).
[31] ***Добров, Б.В***. Онтология по естественным наукам и технологиям ОЕНТ: структура, состав и современное состояние / Б.В. Добров, Н.В. Лукашевич // Электронные библиотеки. – 2008. – Т. 11. – №1.
[32] ***Лукашевич, Н.В***. РуТез-Lite, опубликованная версия тезауруса русского языка РуТез / Н.В. Лукашевич, Б.В. Добров, И.И. Четверкин // Международная конференция по компьютерной лингвистике Диалог-2014. – 2014. - С. 340-349.

DEVELOPING LINGUISTIC ONTOLOGIES IN BROAD DOMAINS

N.V. Loukachevitch[1], B.V. Dobrov[2]

Research Computing Center of Lomonosov Moscow State University, Moscow, Russia
[1] *louk_nat@mail.ru,* [2] *dobrov_bv@mail.ru*

Abstract

The paper describes a model of a linguistic ontology for automatic document processing in a broad domain, that is a domain comprising thousands of entity classes and unrestricted number of possible relations between them. We present a novel set of relations between concepts, which was specially developed for automated document processing and information retrieval. We propose to use a small set of relations comparable with relations in traditional information-retrieval thesauri. However, we use stricter, ontologically motivated definitions for establishing relations. The utilized relations describe the most sufficient relationships between entities, can be applied for description of various domains. The set of concepts relations is illustrated with the example from newly published linguistic ontology – thesaurus of Russian language RuThes.

Key words: *linguistic ontology, broad domain, information retrieval, text analytics.*

Acknowledgments

This work is partially supported by Russian Foundation for Humanities, grant No.15-04-12017.

References

[1] ISO 25964-1:2011, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval / Geneva: International Organization for Standards, 2011.
[2] ANSI/NISO Z39.19-2005, Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. – Bethesda, MD: NISO Press, 2005.
[3] ***Miller, G.*** Nouns in WordNet / G. Miller // WordNet – An Electronic Lexical Database. – The MIT Press, 1998. – P. 23-47.
[4] ***Berners-Lee, T.*** The Semantic Web / T. Berners-Lee, J. Handler, O. Lassila // Scientific American - 2001. – V. 284. – No 5. – P. 28-37.
[5] ***Nirenburg, S***. What's in a symbol: Ontology, representation, and language / S. Nirenburg, Y. Wilks // Journal of Experimental and Theoretical Artificial Intelligence. - 2001. - V. 13(1). - P. 9-23.
[6] ***Lande, D.V.*** Podhod k sozdaniju terminologicheskih slovarei [Approach to the creation of terminological ontologies] / D.V. Lande, A.A. Snarskii // Ontologia proektirovania. – 2014. – No. 2(8). – P. 49-55. (In Russian).
[7] ***Sowa, J.*** Building, Sharing and Merging Ontologies. - http://www.jfsowa.com/ontology/ontoshar.htm (Valid on 07.03.2015).

[8] ***Magnini, B***. Merging Global and Specialized Linguistic Ontologies / B. Magnini, M. Speranza // Proceedings of OntoLex. - 2002. - P. 43-48.
[9] ***Veale, T.*** A context-sensitive framework for lexical ontologies / T. Veale, Y. Hao // Knowledge Engineering Review. - 2007. -Vol. 23(1). - P. 101-115.
[10] ***Guarino, N***. Ontologies and Knowledge Bases: Towards a Terminological Clarification / N. Guarino, P. Giaretta // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. - Amsterdam: IOS Press, 1995. - P. 25-32.
[11] ***Kleschev, A.S***. Klassifikacija svoistv ontologii. Ontologii i ih klassifikacii [Classification of ontology properties. Ontologies and their Classification] / A.S. Kleschev, E.A. Shalfeeva // NTI ser. 1. - 2005. – No. 9. - P. 16-22. (In Russian).
[12] ***Corcho, O***. Roadmap to Ontology Specification Languages / O. Corcho, A. Gomez-Perez // Knowledge Engineering and Knowledge Management. Methods, Models and Tools. / Eds: R. Dieng and O. Corby. - Springer, 2000. - P. 80-96.
[13] ***Gavrilova, T.A.*** Bazy znanij intellektual'nyh sistem [Knowledge Bases on Intellectual Systems] / T.A. Gavrilova, V.F. Horoshevskij. – SPb.: Piter, 2001. – 384 p. (In Russian).
[14] ***Maedche, A***. Learning Ontologies for the Semantic Web / A. Maedche, S. Staab // Proceedings of Semantic Web Workshop. – Hongkong, 2001.
[15] ***Buitelaar, P***. Towards Linguistically Grounded Ontologies. The Semantic Web: Research and Applications / P. Buitelaar, Ph. Cimiano, P. Haase, M. Sintek // Proceedings of the European Semantic Web Conference. LNCS-5554. - Springer Verlag, 2009. - P. 111-125.
[16] GOST 7.25.-2001. Thesaurus for Information Retrieval: Guidelines of developing: structure, and form of representation // System of Standards on Information. – Minsk: Interstate council on Standardization, metrology and certification. 2001. (In Russian).
[17] ***Guarino, N***. Some Ontological Principles for Designing Upper Level Lexical Resources / N. Guarino // Proceedings of First International Conference on Language Resources and Evaluation. - Granada, Spain, 1998.
[18] ***Gomez-Perez, A***. OntoWeb. Technical Roadmap. D.1.1.2. / A. Gomez-Perez, M. Fernandez-Lopez, O. Corcho // IST project IST-2000-29243, 2001.
[19] ***Sowa, J***. Knowledge Representation: Logical, Philosophical, and Computational Foundations / J. Sowa // Brooks Cole Publishing Co., Pacific Grove, CA. 2000.
[20] ***Guarino, N.*** Evaluating ontological decisions with ONTOCLEAN / N. Guarino, C. Welty // Communications of the ACM. - 2002. - V. 45(2). - P. 61-65.
[21] ***Masolo, C.***. Social roles and their descriptions / C. Masolo**,** L. Vieu, E. Bottazzi, C. Catenacci, R. Ferrario, A. Gangemi, N. Guarino // Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning. - AAAI Press. 2004.
[22] ***Steinmann, F***. The representation of roles in object-oriented and conceptual modelling / F. Steinmann // Data and Knowledge engineering. - 2000. - V. 35. – No. 1. - P. 83-106.
[23] ***Winston, M.*** A Taxonomy of Part-Whole Relations / M. Winston, R. Chaffin, D. Hermann // Cognitive Science/ - 1989. - No. 11. - P. 417-444.
[24] ***Niles, I***. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology / I. Niles, A. Pease // Proceedings of the IEEE International Conference on Information and Knowledge Engineering.- 2003. - P. 412-416.
[25] ***Masolo, C.*** WonderWeb. Final Report / C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, L. Shneider // Deliverable D18. 2003.
[26] ***Loukachevitch, N.V.*** Tezaurusy v zadachah informacionnogo poiska [Thesauri in Information-Retrieval Tasks] / N.V. Loukachevitch / - Moscow: Publishing House of Moscow State University. 2011. – 512 p. (In Russian).
[27] ***Guizzardi, G.*** Ontological Foundations for Conceptual Part-Wholes Relation: The Case of Collectives and Their Parts / G. Guizzardi // Advanced Information Systems Engineering. Springer CAiSE. LNCS 6741. – Springer, 2011. - P. 138–153.
[28] ***Loebe, F***. Abstract vs. Social Roles: A Refined Top-level Ontological Analysis / F. Loebe // Proceedings of the 2005 AAAI Fall Symposium 'Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems. - AAAI Press, 2005. - P.93-100.
[29] ***Guarino, N***. The ontological Level: Revisiting 30 yers of Knowledge Representation / N. Guarino // Conceptual Modeling: Foundations and Applications. - Springer-Verlag Berlin, Heidelberg, 2009. - P. 52-67.
[30] ***Kumar, A***. The ontology of blood pressure: a case study in creating ontological partitions in biomedicine / A. Kumar, B. Smith. - http://ontology.buffalo.edu/medo/BPO.pdf (Valid on 07.03.2015).
[31] ***Dobrov, B.V.*** Ontologija po estestvennym naukam i technologijam [Ontology on Natural Sciences and Technologies: Structure and Current State] / B.V. Dobrov, N.V. Loukachevitch // Elektronnye biblioteki. – 2008. – V. 11. – No. 1 (in Russian).

[32] ***Loukachevitch, N.V.*** RuTez-Lite, opublikovannaja versija tezaurusa russkogo jazyka [RuThes-Lite, published version of thesaurus of Russian language RuThes] / N.V. Loukachevitch, B.V. Dobrov, I.I. Chetviorkin // Proc. of international conference on computational linguistics Dialog-2014. – 2014. - P. 340-349.

Сведения об авторах

Лукашевич Наталья Валентиновна, 1964 г. рождения. Окончила факультет Вычислительной математики и кибернетики МГУ им. М.В. Ломоносова в 1986 г., к.ф.-м..н. (1989). Ведущий научный сотрудник НИВЦ МГУ им. М.В. Ломоносова. В списке научных трудов более 150 работ в области автоматической обработки текстов, представления знаний.

Loukachevitch Natalia Valentinovna (b. 1964) graduated from Lomonosov Moscow State University in 1986, PhD (1989). She is leading researcher in Research Computing Center of Lomonosov Moscow State University. She is author of more 150 scientific papers in the field of natural language processing.

Добров Борис Викторович, 1963 г. рождения. Окончил факультет Вычислительной математики и кибернетики МГУ им. М.В. Ломоносова в 1985 г., к.ф.-м..н. (1988). Заведующий лабораторией НИВЦ МГУ им. М.В. Ломоносова. В списке научных трудов более 100 работ в области информационного поиска, онтологий

Dobrov Boris Vikotorovich (b. 1963) graduated from Lomonosov Moscow State University in 1985, PhD (1988). Chief of laboratory Research Computing Center of Lomonosov Moscow State University. He is author of more than 100 publications in the field of information retrieval, ontologies.