УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 149, кн. 2 Физико-математические науки *2007*

УДК 519.765

ЛИНГВИСТИЧЕСКАЯ ОНТОЛОГИЯ ПО ЕСТЕСТВЕННЫМ НАУКАМ И ТЕХНОЛОГИЯМ ДЛЯ ПРИЛОЖЕНИЙ В СФЕРЕ ИНФОРМАЦИОННОГО ПОИСКА

Б.В. Добров, Н.В. Лукашевич

Аннотация

В статье описывается идея, методология и текущее состояние проекта по созданию лингвистической онтологии – специального информационно-поискового тезауруса для автоматической обработки текстов по естественным наукам. В настоящее время ресурс содержит более 30 тыс. понятий, 70 тыс. терминов для таких научных дисциплин, как математика, физика, химия, геология и биология. В статье также рассматриваются типы изменений описаний понятий, происходящих при перемещении описаний из общезначимой лингвистической онтологии в лингвистическую онтологию конкретной прикладной области.

Введение

Эффективное решение задач информационного поиска научно-технической информации является одним из условий перехода отраслей экономики на качественно новые технологические уровни.

Большое распространение получили глобальные машины поиска, обеспечивающие поиск на основе лексического совпадения запроса и документа. Для профессионального, в том числе научно-технического, поиска информации требуется обеспечение поиска, основанного на знаниях, – использование синонимов, возможности автоматического расширения запроса, возможностей автоматического анализа результатов запроса и помощь в интерактивном поиске.

Традиционными средствами тематического поиска научной информации в течение многих лет являлись информационно-поисковые тезаурусы. Однако такие тезаурусы создавались для их использования в процессе ручного индексирования и поиска, и не обеспечивают эффективного информационного поиска в автоматических режимах [1, 2].

В настоящее время перспективы организации более качественного, содержательного информационного поиска в сети интернет связываются с разработкой онтологий.

Согласно [3], под онтологиями понимают систему явной концептуализации предметной области, то есть формального представления предметной области.

Отметим, что существуют разные формальные интерпретации [4–7] столь нечеткого определения. Общим для всех формализаций является выделение множества объектов (концептов, понятий), алфавита отношений, правил установления отношений и аксиом, задающих правила вывода на множестве отношений.

С точки зрения использования онтологий в задачах автоматической обработки текста существует два подхода к установлению соответствия между онтологией предметной области и языком предметной области (лексиконом).

С одной стороны, сначала строится система понятий, которым затем приписываются наборы языковых выражений (слов, терминов, словосочетаний). Обнаружение этих выражений в тексте позволяет инициировать соответствующие понятия и связанные с ними правила [3].

С другой стороны, замечено, что существующие лингвистические ресурсы (словари, глоссарии, тезаурусы) также задают определенную концептуализацию предметной области.

В результате, согласно современным воззрениям, термину «онтология» удовлетворяет широкий спектр структур, представляющих знания о той или иной предметной области. В качестве в разной степени формализованных онтологий рассматриваются [8]:

1) словарь с определениями;
2) простая таксономия;
3) тезаурус (таксономия с терминами);
4) модель с произвольным набором отношений;
5) таксономия и произвольный набор отношений;
6) полностью аксиоматизированная теория.

Одновременно подчеркивается [9] различие между фундаментальными онтологиями (fundamental ontologies), которые описывают предметную область максимально полно (п. 6), безотносительно к приложениям и обычно с максимальной степенью формализации, и прикладными онтологиями (application ontologies), которые также называются «легкими» онтологиями (lightweight ontologies) и которые формализуются настолько, насколько это необходимо для приложения (пп. 1–5).

Понятно, что создать фундаментальную онтологию для большой области научного знания не представляется возможным в силу существования различных теорий и постоянного изменения трактовки самых базовых понятий.

С другой стороны, степень формализации описания предметных областей в традиционных информационно-поисковых тезаурусах оказалась недостаточной для автоматического расширения запросов в информационном поиске.

Возникает вопрос, какова же должна быть степень формализации предметной области, структура онтологии, чтобы:

– с одной стороны, эту онтологию можно было создать и начать использовать в разумные сроки (2–3 года) относительно небольшим коллективом;

– с другой стороны, чтобы степень формализации понятийной структуры предметной области обеспечивала возможность содержательного информационного поиска в автоматических режимах.

Как уже указывалось, понятия онтологии, предназначенные для поддержки решения задач информационного поиска, должны быть аккуратно связаны со значениями терминов предметной области. Такого рода онтологии называются лингвистическими онтологиями: главной характеристикой лингвистических онтологий является то, что они связаны со значениями (“are bound to the semantics”) языковых выражений (слов, именных групп и т. п.) [9].

В качестве примера лингвистической онтологии часто приводится ресурс WordNet [10]. Этот ресурс представляет в виде иерархической структуры систему значений слов общезначимого английского языка. Вместе с тем возникает достаточно много проектов, которые на основе модели WordNet описывают терминологические системы конкретных предметных областей, то есть создают лингвистические онтологии этих областей [11–13].

Под руководством авторов в 2004 г. были начаты работы над созданием лингвистической онтологии для автоматической обработки в области естественных наук.

В настоящей статье, обобщающей работы [14–16], описывается идея, методология и текущее состояние проекта. В первом разделе рассматриваются достоинства и недостатки существующих лингвистических ресурсов с точки зрения применимости для автоматической обработки научных текстов в сфере естественных наук. Во втором разделе мы описываем идею и основные положения проекта создания лингвистической онтологии для естественных наук. В следующем разделе описывается ранее созданные авторами онтологии – Тезаурус русского языка РуТез, Общественно-политический тезаурус, на основе методологии создания которых создается новая лингвистическая Онтология по естественным наукам и технологиям. В четвертом разделе излагаются этапы построения указанной онтологии, приводятся данные о текущем состоянии ресурса, о способах тестирования создаваемой онтологии. Пятый раздел описывает типы изменений в описаниях понятий, полученных Онтологией по естественным наукам и технологиям» из Тезауруса РуТез.

1. Ресурсы для смыслового анализа электронных коллекций

1.1. Традиционные информационно-поисковые тезаурусы. Хронологически первыми ресурсами, служащими для упорядочения работы с электронными коллекциями, были информационно-поисковые тезаурусы (ИПТ) [17–20], в которых синонимичные термины были собраны вокруг наиболее представительного термина (предпочтительного термина), называемого дескриптором, а между дескрипторами устанавливались отношения.

Однако традиционные информационно-поисковые тезаурусы разрабатывались для ручного индексирования человеком-индексатором, а в настоящее время объем потоков информации значительно превосходит возможности индексаторов по их тематической обработке. Применение традиционных информационно-поисковых тезаурусов при автоматическом индексировании и автоматическом расширении запроса приводит лишь к ухудшению характеристик поиска по сравнению с поиском по словам [1, 21].

Это связано с тем, что традиционный информационно-поисковый тезаурус описывает, по сути дела, искусственный язык, служащий для фиксации основной темы документа. Человек-индексатор должен был перевести естественный язык документа на искусственный язык тезауруса. Поэтому вся процедура разработки и использования информационно-поисковых тезаурусов основывалась на лингвистических и предметных знаниях эксперта. Многие решения, принимаемые в процессе создания тезаурусов, были направлены на то, чтобы сделать работу индексатора более удобной и менее субъективной.

Для того чтобы использоваться в автоматическом режиме, традиционным тезаурусам недостает значительного объема информации:

– описания большого количества понятий более низкого уровня иерархии, чем представленные дескрипторы;

– намного более подробное описание синонимии терминов;

– описания многозначности слов;

– недостаточна также система традиционных отношений между дескрипторами тезауруса и их свойствами, базирующаяся в основном на использовании отношений ВЫШЕ-НИЖЕ и АССОЦИАЦИЯ.

В России наиболее известен Тезаурус научно-технических терминов [18], который издан в 1972 году. Тезаурус описывает терминологию военно-промышленного комплекса 1970-х годов, не соответствует реалиям и технологиям настоящего времени. ВИНИТИ обладает громадным массивом научно-технических текстов, имеются наборы терминов [22] по научно-техническим отраслям. Но эти термины не

организованы иерархическими связями в единый ресурс научно-технической терминологии.

1.2. От информационно-поисковых тезаурусов к фундаментальным онтологиям. Некоторые авторы [2, 23], решая проблему модификации традиционных информационно-поисковых тезаурусов к современным задачам автоматической обработки больших текстовых коллекций, предлагают преобразовать систему отношений тезауруса в более формализованный набор предикатов (уровень формализации 5, см. введение) и описать правила вывода (аксиомы).

Так, например, в работе [2] в качестве примеров модификации информационно-поискового тезауруса по сельскому хозяйству AGROVOC приводятся следующие словарные статьи:

Исходные статьи тезауруса AGROVOC (NT – отношение НИЖЕ, BT – отношение ВЫШЕ):

milk
NT cow milk
NT milk fat

cow
NT cow milk

Cheddar cheese
BT cow milk

Преобразованные словарные статьи выглядят следующим образом:

milk
<*includesSpecific*> cow milk
<*containsSubstance*> milk fat

cow
<*hasComponent*> cow milk

Cheddar cheese
<*madeFrom*> cow milk

Пример предлагаемых правил вывода:

Правило 1:
Part_X <*mayContainSubstance*> Substance_Y
IF Animal_W <*hasComponent*> Part_X
AND Animal_W <*ingests*> Substance_Y

Правило 2:
Food_Z <*containsSubstance*> Substance_Y**:**
IF Food_Z <*madeFrom*> Part_X
AND Part_X <*containsSubstance*> Substance_Y

Предполагается, что система, имея такие правила вывода, может автоматически получить, что сыр-чеддер содержит (<*containsSubstance*>) молочный жир, и, что если коровы на ферме съели корма, зараженные ртутью, то, сыр, сделанный из этого молока, также, возможно, будет заражен ртутью (*Cheddar cheese* <*mayContainSubstance*> *mercury*).

Однако, чтобы такой вывод действительно был сделан, нужно, помимо изменений в описании понятий и терминов предметной области, иметь автоматические

средства обработки естественно-языковых текстов, позволяющие в неограниченном связном тексте точно и полно извлекать последовательности фактов, уметь прослеживать кореферентность, следить за временем извлекаемых фактов: в корма попала ртуть, эти корма принадлежат данной ферме, коровы этой фермы съели именно эти корма, изготовление сыра чеддер этой фермой произведено в период времени сразу после того, как эти коровы съели эти корма и т. п.

Кроме того, в тексте слова *корма* и *ртуть* могут оказаться в разных частях длинного предложения, или в разных предложениях текста, например, из-за использования эллиптической конструкции или местоимения и т. п., что значительно усложнит выявление указанного выше факта.

Понятно, что в настоящее (и ближайшее) время ни одна из существующих систем автоматической обработки текстов, извлечения знаний из текстов не может обеспечить такой уровень точности и полноты получения информации из текстов, на которых можно было бы надежно обосновывать работу таких правил вывода.

Таким образом, по нашему мнению, значительные трудозатраты на такого рода формализацию информационно-поисковых тезаурусов не приведут к улучшению качества автоматической обработки текстов и созданию ресурсов, лучше приспособленных к автоматическим режимам работы, чем существующие информационно-поисковые тезаурусы.

1.3. Отношения в онтологии, применяемой в неопределенных контекстах. На основе анализа, проведенного в предыдущем пункте, можно заметить, что информационно-поисковые онтологии в течение долгого времени будут вынужденно применяться в условиях неопределенного контекста, то есть в условиях, когда ни об одном выявленном в тексте понятии не будет точно и полно известен даже набор явно упоминаемых о нем в тексте фактов и других видов информации. Таким образом, в таких условиях надежно могут использоваться лишь отношения, которые не зависят или слабо зависят от конкретного текста, то есть те, которые не исчезают, не меняются в течение всего срока существования любого или подавляющего большинства экземпляров понятия. Например, «любой лес всегда состоит из деревьев».

Наиболее известным типом отношения, которое выполняется для всех экземпляров, является таксономическое отношение. Так, если $C1$ упомянуто в тексте и $C1$ является видом $C2$, это означает, что в тексте упомянуто и $C2$. Если данный текст релевантен запросу о $C1$, то он будет релевантен и запросу о $C2$.

В условиях, когда невозможно использование сложных правил вывода, для осуществления вывода по тексту желательно найти другие типы отношений, обладающих свойствами транзитивности и наследования подобно таксономическим отношениям.

Как нам представляется, именно такого рода отношениями являются отношения онтологической зависимости, изучаемые в рамках философской дисциплины «формальная онтология» [4, 5].

Отношения онтологической зависимости описывают, подразумевает ли существование одного понятия существование каких-либо других понятий. Эти отношения подразделяются на следующие виды:

– подразумевает ли существование сущности существование чего-либо еще (строгая зависимость – rigid dependence), например, *кипение* невозможно без существования конкретного объема жидкости, которая кипит;

– предполагается ли существование примеров некоторого класса (родовая зависимость – generic dependence) некоторых сущностей, например, возникновение

понятия *гараж* невозможно без существования понятия *автомобиль*, хотя конкретный гараж может возникнуть безотносительно к конкретному автомобилю;

– предполагает ли существование X в некоторый момент времени T существование Y в некоторый другой момент времени $T1$ (историческая зависимость), например, *солома* исторически зависит от *молотьбы*, поскольку *солома* не может возникнуть без предварительного процесса *молотьбы*, вместе с тем, когда этот процесс заканчивается, солома может длительное время продолжать существовать.

В работе [24] постулируется транзитивность отношений онтологической зависимости.

В работах [25, 26] было показано, что отношения строгой и родовой онтологической зависимости эффективны при создании ресурсов для информационного поиска.

1.4. WordNet как лингвистическая онтология. Целью разработки WordNet [10] являлось не описание системы понятий, а установление системы отношений между лексическими значениями.

Между значениями слов и понятиями имеется достаточно сложная взаимосвязь: «значение шире понятия, так как включает в себя оценочный и ряд других компонентов, значение у́же понятия в том смысле, что включает лишь различительные черты объектов, а понятия охватывают их наиболее глубокие существенные свойства. . . »[27].

Наиболее ярко различие между описаниями лексики и иерархии понятий в ресурсах типа WordNet проявляется в расчленении иерархической сети на подсети по частям речи, когда совпадающим по значению, но различающимся по частям речи словам (например, *приватизация*, *приватизировать*, *приватизационный*) соответствуют разные узлы иерархической сети. Ясно, что понятие, соответствующее этим словам, должно быть одно и то же.

Многие типы отношений в ресурсах класса WordNet, такие как отношения *антоним*, *дериват*, *валентности* [28], описывают отношения между лексическими единицами, а не понятиями.

В конкретных предметных областях значения предметной лексики и понятия предметной области максимально сближаются, но применяемые при разработке WordNet-подобных ресурсов в конкретных предметных областях методы (модели, отношения) остаются теми же, что и для описания общезначимой лексики.

При создании WordNet-подобных ресурсов в конкретных предметных областях роль концептуального анализа понятийной модели предметной области играет меньшую роль по сравнению с информационно-поисковыми тезаурусами, при разработке которых связь «термин-понятие» предметной области осознавалась достаточно четко.

В то же время внимание разработчиков WordNet-подобных ресурсов в конкретных предметных областях к каждой языковой единице, работа со значениями предметной лексики являются необходимыми для автоматизации обработки предметных текстов, поскольку путь к понятийному содержанию того или иного текста лежит через совокупность конкретных языковых выражений этого текста.

Итак, подчеркнем, что в информационно-поисковых тезаурусах недостаточно представлена связь понятий предметной области с лексикой конкретных текстов, в WordNet-подобных ресурсах ослаблена понятийная сторона описания предметной лексики. Между тем, для успешного автоматического анализа предметно-ориентированных текстов описание «понятие – язык предметной области» должно быть сбалансировано: описание предметной лексики невозможно без анализа

понятийной модели предметной области, распознавание понятийного содержания текстов невозможно без качественного описания языка предметной области.

Лингвистической онтологией, в которой была сделана попытка такого сбалансированного подхода к описанию системы значений языковых единиц и связанной с ними системы понятий, является онтология Mikrokosmos [29].

2. Проект разработки новой лингвистической онтологии

Проект предлагает создание лингвистической онтологии для обеспечения автоматической обработки научно-технической информации – понятийного индексирования, автоматической классификации потока научно-технической информации.

Создаваемая лингвистическая онтология строится на сочетании трех различных традиций и методологий:

1) методологии разработки информационно-поисковых тезаурусов;

2) методологии разработки лингвистических ресурсов типа WordNet (Принстонский университет);

3) методологии созданий формальных онтологий.

В методологии разработки информационно-поисковых тезаурусов важны следующие принципы:

– единицы тезауруса создаются на основе терминологии;

– описание большого числа многословных выражений, принципы включения (невключения) многословных единиц;

– простой набор отношений между единицами.

В методологии разработки лексических ресурсов типа WordNet важны следующие положения:

– многоступенчатое иерархическое построение лексико-терминологической системы понятий;

– технология описания значений многозначных слов и выражений.

В методологии разработки формальных онтологий важными являются:

– разработка лингвистической онтологии как иерархической системы понятий;

– строгость построения таксономии, отличие истинно таксономических отношений от ролевых отношений;

– использование для описания нетаксономических отношений онтологической зависимости;

– использование в качестве аксиом (правил вывода) свойств транзитивности и наследования таксономических отношений и отношений онтологической зависимости.

Разработка новой лингвистической онтологии состоит из следующих этапов.

Прежде всего создается большой корпус текстов, принадлежащий предметной области, для которой создается онтология.

С помощью разного рода автоматизированных процедур из текста извлекаются значимые в предметной области слова и словосочетания.

После этого с корпусом, а также со словарями предметной области начинают работать эксперты.

Основными целями их работы являются следующие.

– Изучая конкретные языковые выражения, их словарные определения, употребление в конкретных текстах определить, какому понятию соответствует значение данного языкового выражения. Если такое понятие уже существует, данное языковое выражение приписывается этому понятию. Для нового понятия создается отдельная единица в иерархической сети.

– Для каждого понятия по корпусу набирается максимально возможное число различных слов, выражений, значения которых соответствуют этому понятию. Такие языковые выражения называются текстовыми входами понятия или терминами онтологии.

– Для каждого понятия проводится концептуальный анализ для выяснения его таксономических отношений и отношений онтологической зависимости. Поскольку эти отношения являются наиболее важными для широкого круга понятий, их часто можно выявить на основе анализа определений соответствующих терминов в терминологических словарях, употреблений в текстовых контекстах, сопоставления определений и текстовых контекстов.

Как показывает практика, в связи с многократно описанными проблемами получения знания от экспертов в предметной области [30], наиболее эффективным является максимально полная разработка ресурса на основе анализа текстового корпуса. Далее созданный проект ресурса предъявляется экспертам в предметной области, которые уже достаточно легко находят в нем возможные ошибки и неточности, могут объяснить, почему им не понравилось то или иное отношение.

Следует отметить, что на этапе разработки онтологии в качестве экспертов выступают лингвисты, которые имеют опыт работы с текстовыми корпусами, лексическими значениями. Помимо авторов, в разработке онтологии принимали участие эксперты-лингвисты О.А. Штернова, Т.М. Селиванова, И.А. Каргина.

Основная парадигма авторов проекта состоит в том, что базисом для автоматического смыслового анализа текстов, в том числе для Semantic Web, должны действительно стать онтологии предметных областей, но это должны быть БОЛЬШИЕ онтологии, ориентированные на основную среду обмена информации – текстовую информацию.

В самом деле, подробные сетки понятий, описываемые с единых понятных всем «языковых» позиций, должны обеспечивать возможность интеграции онтологий разных предметных областей по пересекающимся понятиям.

Данный вывод авторы проекта делают на основе имеющегося опыта создания больших лингвистических онтологий для нескольких предметных областей: области общественно-политических отношений (лексика правовых документов и материалов СМИ), области технической авиационной документации, области спецификаций на программное обеспечение, области компьютерной безопасности.

3. Отправная точка

3.1. Ранее созданные ресурсы. Авторы проекта ранее [31, 32] создали информационно-поисковый тезаурус для автоматического индексирования текстов в общественно-политической области (далее – Общественно-политический тезаурус), включающих более 32 тыс. понятий, 79 тыс. русскоязычных и 80 тыс. англоязычных текстовых входов.

Представляя собой по форме информационно-поисковый тезаурус с ограниченным набором отношений, Общественно-политический тезаурус построен на основе формальных онтологических принципов. Это позволяет нам позиционировать его как лингвистическую онтологию для автоматической обработки документов в области общественно-политических отношений.

Создан [25, 33] не только лингвистический ресурс, но и комплекс математического обеспечения (моделей, алгоритмов) и программного обеспечения (утилит, информационных систем), то есть создан полный технологический цикл от набора терминологии до реализации обеспечения функционирования информационно-аналитических систем различного назначения.

Общественно-политический тезаурус используется как лингвистический ресурс в таких задачах информационного поиска, как автоматическое концептуальное индексирование, визуализация результатов поиска, автоматическая рубрикация документов, автоматическое аннотирование.

С 1998 г. Общественно-политический тезаурус вошел в состав Тезауруса русского языка РуТез, который теперь, помимо общественно-политической терминологии, содержит описание значений широкого круга общезначимой лексики в виде сети понятий и поэтому также рассматривается нами как лингвистическая онтология. Далее мы будем ссылаться на лингвистическую онтологию Тезаурус РуТез, подразумевая в его составе Общественно-политический тезаурус.

Для реализации обсуждаемого проекта наиболее важны созданные ранее технологии быстрого автоматизированного формирования [34] терминологической базы по текстам, а также возможность использования уже существующего ресурса большого объема.

В общественно-политических текстах понятия общественных наук встречаются значительно чаще, чем понятия естественных наук, что находит свое отражение в составе Общественно-политического тезауруса. Тем не менее сфера естественных наук затрагивается в связи с обсуждением вопросов промышленности, нефтедобычи, медицины и т. п., поэтому соответствующие научная лексика и терминология неплохо представлены в тезаурусе, что позволило поставить задачу их использования при создании нового ресурса.

3.2. Причины раздельного ведения онтологий. Начало работ над Онтологией по естественным наукам и технологиям означало, что было принято решение раздельно разрабатывать две разные онтологии для анализа текстов в общественно-политической сфере (газетные статьи, новостные сообщения, законодательные акты, международные договоры) и научных публикаций.

Решение о разделение онтологий было связано с несколькими серьезными факторами.

Во-первых, обе онтологии достаточно объемны, включают десятки тысяч понятий и отношений, при этом большая часть понятий общей онтологии обычно не используется в текстах естественных наук, и, наоборот, научные понятия по большей мере не нужны для анализа таких общезначимых документов, как газетные статьи, информационные сообщения, законодательные акты.

Во-вторых, разделение онтологий снижает многозначность описанных слов и выражений.

В-третьих, предполагалось, что существует несоответствие так называемой «бытовой» картины мира и научной картины мира, то есть отношения, описанные и правильные в рамках одной онтологии, должны быть изменены в рамках другой онтологии.

И наконец, последнее (по перечислению, но не по важности), эти две онтологии отличаются по способам рассмотрения внешнего мира: онтология РуТез рассматривает мир через призму современного цивилизованного общества – что известно о мире значимому количеству образованных людей современного общества, что важно (воздействует, используется) в существовании современного общества. Онтология в области естественных наук и технологий исключает из рассмотрения аспекты общественного мировосприятия и должна описывать в виде онтологической модели устоявшиеся воззрения современной науки на основе материалов научных публикаций.

Вместе с тем хотелось бы отметить, что существуют типы текстов, для анализа которых могут понадобиться обе онтологии, работающие одновременно, и поэто-

му нужно иметь четкое представление об отражении сходных явлений в разных контекстах.

К числу текстов, требующих, как нам представляется, использования обеих онтологий, относятся:

– анализ соответствий между требованиями технического регулирования и описанием производственных процессов;

– документы вида «заявки/отчеты» о научном исследовании;

– инвестиционные заявки, связанные с промышленным внедрением научных исследований.

3.3. Структура онтологий. Оба ресурса – Тезаурус РуТез и Онтология по естественным наукам и технологиям – имеют одинаковую структуру. Они являются онтологиями, поскольку описывают понятия внешнего мира и отношения между ними, которые устанавливаются в соответствии с требованием правомочности расширения запроса по иерархии связей при информационном поиске. Оба ресурса принадлежат к особому классу онтологий, так называемым лингвистическим онтологиям [8, 35], поскольку введение понятий в значительной мере мотивируется значениями языковых единиц, относящихся к предметной области ресурса. Далее в этом разделе мы опишем структуру этих ресурсов, ссылаясь на них обобщенным названием РуТез*Онтологии.

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.

Опишем сказанное более подробно.

РуТез*Онтология – это иерархическая сеть понятий. Каждое понятие имеет имя.

Для сопоставления с текстом каждое понятие снабжается набором текстовых выражений («*текстовых входов*», «*терминов*»), значения которых соответствуют данному понятию. В качестве таких текстовых входов могут выступать однословные существительные, прилагательные, глаголы, именные и глагольные группы. Количество таких текстовых входов понятий может быть достаточно велико, например, превышать 20 единиц. При вводе нового понятия делаются специальные усилия, чтобы максимально подробно перечислить его возможные текстовые входы.

Каждое понятие связывается отношениями с другими понятиями РуТез*Онтологии. Набор отношений РуТез*Онтологии специально подобран для эффективной работы в информационно-поисковых приложениях.

В РуТез*Онтологии имеется четыре основных типа отношения.

Первый тип отношений – родовидовое отношение НИЖЕ-ВЫШЕ – обладает свойством транзитивности и наследования.

Второе тип отношений – отношение ЧАСТЬ-ЦЕЛОЕ. Используется не только для описания физических частей, но и для других внутренних сущностей понятия, таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жестко связаны со своим целым, то есть каждый пример понятия-части должен в течение всего времени своего существования являться частью для понятия-целого и не относиться к чему-либо другому.

Так, например, в РуТез*Онтологиях считается неправильным описывать понятие *ДВИГАТЕЛЬ* частью понятия *АВТОМОБИЛЬ*, поскольку двигатели являются частями различных технических устройств, а не только автомобилей. Мы

вводим понятие *АВТОМОБИЛЬНЫЙ ДВИГАТЕЛЬ* как видовое понятие для понятия *ДВИГАТЕЛЬ* и затем устанавливаем отношение ЧАСТЬ между понятием *АВТОМОБИЛЬ* и понятием *АВТОМОБИЛЬНЫЙ ДВИГАТЕЛЬ*.

В этих условиях удается выполнить свойство транзитивности введенного таким образом отношения ЧАСТЬ-ЦЕЛОЕ, что очень важно для автоматического вывода в процессе автоматической обработки текстов.

Еще один тип отношения, называемого несимметричной ассоциацией АСЦ2-АСЦ1, связывает два понятия, которые не могут быть связаны выше рассмотренными отношениями, но одно из понятий не существовало бы без существования другого. Например, понятие *ЛЕС* требует существования понятия *ДЕРЕВО* (при этом можно было бы ввести как ЧАСТЬ для понятия *ЛЕС* понятие *ДЕРЕВО В ЛЕСУ*), а понятие *АНТИСТАТИК* требует существования понятия СТАТИЧЕСКОЕ *ЭЛЕКТРИЧЕСТВО*.

Последний тип отношений – симметричная ассоциация – связывает, например, понятия, очень близкие по смыслу, но которые мы не решились «склеить» в одно понятие.

Отношения НИЖЕ-ВЫШЕ, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация являются иерархическими отношениями. Таким образом, на основе свойств иерархичности, транзитивности и наследования для каждого понятия может быть определена совокупность понятий, которые являются для него нижестоящими понятиями по иерархии.

4. Этапы реализации проекта

Основной задачей при создании лингвистической онтологии большого размера силами небольшого коллектива является максимальное использование методов автоматизации, а также фрагментов ранее созданных лингвистических онтологий.

4.1. Автоматический набор терминологии по текстам. Для каждой науки из рассматриваемого списка (математика, физика, химия, биология, геология) были сформированы коллекции документов (от 3000 до 8000 документов, от 50 до 90 Мб). Источником коллекций являлись документы, доступные в интернет, следующих основных типов:

– материалы школьных уроков;
– рефераты;
– университетские лекции;
– материалы специализированных сайтов.

Была произведена обработка специальными процедурами автоматического извлечения терминоподобных словосочетаний, что дало возможность проверки употребимости терминов в материалах, а также нахождения терминов, входящих в состав предметной области.

Для выявления терминов было проведено сопоставление с терминами Общественно-политического тезауруса. Также были применены два алгоритма выделения терминоподобных слов и словосочетаний [34].

Первый алгоритм выделяет существительные, прилагательные, согласованные пары и тройки прилагательных и существительных, а также генеративные конструкции (существительное + существительное в родительном падеже и т. п.).

Второй алгоритм может выделять часто повторяющиеся именные группы в несколько слов, в том числе предложные.

При этом многословные термины, словосочетания из тезауруса РуТез могли выступать «зародышами» для формирования более длинных словосочетаний.

Полученные терминоподобные слова и словосочетания упорядочивались по убыванию суммарной частотности и убыванию количества содержащих их документов.

4.2. Автоматизированное формирование первой версии онтологии. Основной целью при формировании первой версии ресурса являлось быстрое получение приближения предметной области. При этом выбор делался в сторону большей избыточности первого приближения, чтобы в дальнейшем минимизировать по возможности поиск и добавление новых терминов.

4.2.1. Отбор новой терминологии. По каждой предметной области были образованы верхние части частотных списков терминоподобных слов (по 10 тыс.) и словосочетаний (по 15 тыс.), которые были направлены на быструю разметку экспертам. Отметим, что нижняя часть списков соответствовала уровню встречаемости в 5–6 документах.

Эксперты должны были в рамках «своей» науки пометить принадлежность к предметной области того или иного термина. Допускалась пометка термина для нескольких предметных областей, но полнота такого рода разметки не требовалась. После окончания этого этапа списки разных экспертов были объединены – получился список из 32 тыс. помеченных слов и словосочетаний.

4.2.2. Использование существующего ресурса. Существующий ресурс – Общественно-политический тезаурус – покрывает лексику и терминологию нормативно-правовых актов и материалов СМИ, поэтому имеет значительное пересечение с терминологией практически любой значимой предметной области.

Для каждой новой предметной области было задано несколько понятий верхнего уровня, таких как =*НАУКА*=, =*РАСТЕНИЕ*= и т. п., касающихся сущности исследуемых предметных областей и их предметов ведения. Для таких понятий были выбраны способы расширения по иерархии тезаурусных связей (полное расширение или расширение только по таксономическим отношениям). Полученные группы понятий были обозначены специальными пометками отнесения к дополнительной предметной области соответствующей науки и к специальной служебной рабочей предметной области «кандидат».

4.2.3. Пересечение отобранных терминов и существующего ресурса. Список отобранных экспертами терминов по текстам был сопоставлен с текстовыми входами понятий Общественно-политического тезауруса. В случае совпадения с текстовым входом из тезауруса все понятия, ассоциированные с данным текстовым входом, получали дополнительные пометки новых предметных областей – соответствующей науки (наук) и предметной области «кандидат».

Если отобранный экспертами термин был неизвестен, то заводилось новое понятие, дескриптор и единственный текстовый вход которого совпадали с данным термином. Новое понятие получало пометки принадлежности к предметной области соответствующей науки и «кандидат». Кроме того, автоматически вводилось таксономическое отношение ВЫШЕ к специальному временному понятию в каждой науке, например, =@*ГЕОЛОГИЧЕСКАЯ ТЕРМИНОЛОГИЯ*=, =@*ХИМИЧЕСКАЯ ТЕРМИНОЛОГИЯ*= и т. п.

4.2.4. Замыкание предметной области. Для отобранных из тезауруса РуТез понятий (получивших пометку «кандидат») было выполнено «замыкание»: были добавлены понятия, расположенные выше по таксономическим связям. Эти понятия получали аналогичные дополнительные пометки предметных областей.

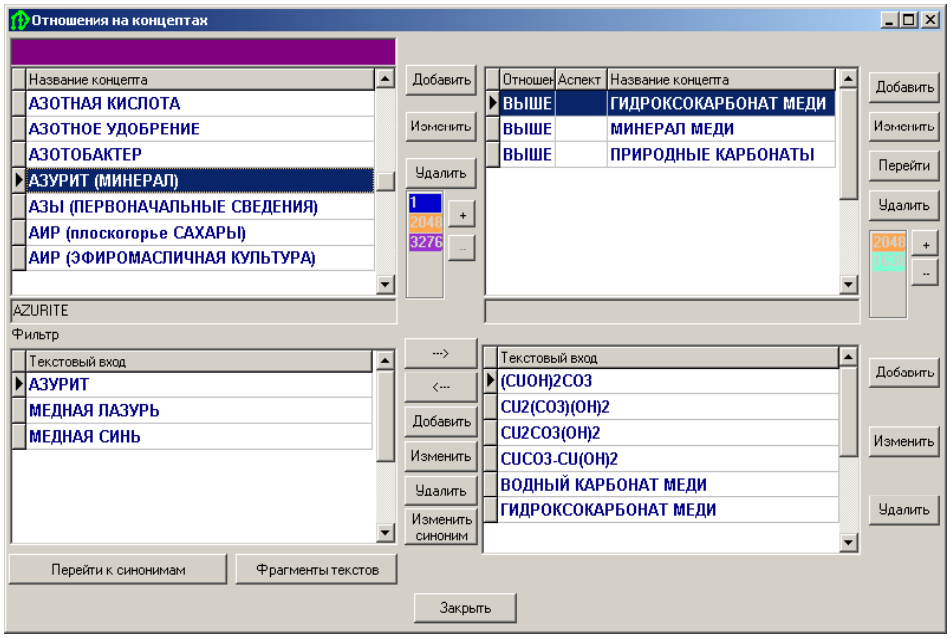

Рис. 1. Основная экранная форма редактирования отношений и текстовых входов понятий

4.2.5. Оформление первой версии ресурса. В результате предыдущих этапов был сформирован «пополненный» ресурс на основе Общественно-политического тезауруса. Так как все интересующие нас понятия имели пометку отнесения к служебной предметной области «кандидат», то мы использовали стандартную процедуру «экспорта» фрагмента тезауруса для формирования нового ресурса.

4.3. Методология работы экспертов. Каждый эксперт может выбрать список понятий, имеющих пометку соответствующей предметной области. Кроме того, эксперт просматривает понятия, связанные отношением с временным служебным понятием типа =@*ХИМИЧЕСКАЯ ТЕРМИНОЛОГИЯ*= (см. п. 4.2.3).

Цель работы эксперта:

– снять пометку «кандидат» с понятий, которые действительно относятся к предметной области соответствующей научной дисциплины;

– снять ложно поставленные пометки принадлежности понятия к предметной области, оставив только пометку «кандидат», либо удалить такое понятие;

– сделать так, чтобы не осталось понятий, подчиненных временному понятию типа =@*ХИМИЧЕСКАЯ ТЕРМИНОЛОГИЯ*=. При этом либо понятие получает новые нетривиальные связи, либо объединяется с существующим, передавая ему свои текстовые входы, либо удаляется.

Естественно, эксперт имеет возможность и непосредственного ввода нового понятия.

На рис. 1 представлен рабочий экран системы ведения онтологии. В левом верхнем углу помещены понятия онтологии, в левом нижнем углу представлены текстовые входы для понятия, на котором установлен курсор *АЗУРИТ (МИНЕРАЛ)* – *азурит, медная лазурь, медная синь*. В левой нижней части экранной формы показаны текстовые входы для понятия. В правом верхнем углу показаны отношения этого понятия. Оно описывается как подкласс понятий *КАРБОНАТ МЕДИ, МИНЕРАЛ МЕДИ, ПРИРОДНЫЕ КАРБОНАТЫ*. Курсор установлен на отношении

с понятием *ГИДРОКСОКАРБОНАТ МЕДИ*. Правая нижняя часть экранной формы показывает текстовые входы понятия, выделенного курсором в правой части экранной формы.

Видно, что экран отражает отношения между традиционно геологическими и химическими понятиями. Таким образом, отражение понятий, традиционно относящихся к разным наукам, в рамках единого ресурса дает возможность использовать для описания отношений понятий разных наук.

В настоящее время экспертами используются следующие три основные источника:

– профильные и общие энциклопедии, толковые словари – как источник профессиональной информации;

– накопленные списки терминоподобных слов, словосочетаний, которые очень эффективны при добавлении синонимов, вариативно отличающихся от указанных в опубликованных энциклопедических источниках;

– каждый текстовый вход должен быть проверен экспертом по употреблению в интернете. Такая проверка производится с использованием глобальных поисковых машин.

4.4. Текущее состояние проекта. В настоящее время онтология включает в себя 30 тыс. понятий, 70 тыс. терминов из таких областей науки, как математика, физика, химия, геология, биология. Покрытие химической и биологической терминологии в пределах средней школы полностью завершено. Размер биологической подонтологии будет зафиксирован на достигнутом уровне. Покрытие терминологии математики и физики в пределах школьной программы будет завершено в течение ближайшего года. После окончания работ с терминологией в рамках школьных программ эксперты начинают работать с программами по отраслям естественных наук (аналитическая химия, историческая геология и т. п.), проверяя полноту отражений терминологии в онтологии.

4.5. Тестирование онтологии. Онтология, предназначенная для автоматической обработки текстов, должна прежде всего тестироваться на текстах соответствующих предметных областей.

Разработан специальный интерфейс, позволяющий изучать результаты автоматической обработки текстов на основе онтологии. Программа позволяет проанализировать:

– совокупность терминов, обнаруженных в тексте;

– терминологическую многозначность в тексте: многозначность (однозначность) термина, результаты разрешения многозначности;

– отношения между понятиями, выявленными в тексте;

– тематическую аннотацию текста – совокупность близких по смыслу понятий (тематические узлы), которые наиболее полно характеризуют содержание текста [33]. Просмотр такой аннотации, сопоставление ее с заголовком текста может выявить неправильно описанные отношения между понятиями, пропущенные отношения, неправильно разрешенную многозначность без подробного просмотра текста. Так, например, если в математической статье, посвященной обсуждению *генетических алгоритмов*, автоматически (на основе знаний онтологии) выделился крупный узел близких по смыслу терминов: *методы оптимизации, метод приведенного градиента, метод имитации обжига*, а сами *генетические алгоритмы* в этот узел не вошли, то эта неточность описаний онтологии видна с первого взгляда на тематическую аннотацию, построенную для данного текста;

– результаты автоматической рубрикации текста, могут быть подсвечены совокупности терминов, на основе которых выведена та или иная рубрика.

5. Изменения в описаниях понятий, полученных из Тезауруса РуТез

Возможность вторичного использования однажды разработанных онтологий в других областях или других приложениях является важной проблемой в онтологических исследованиях [36, 37]. Для поддержки процедуры слияния онтологий и создания на этой основе новой онтологии разработано несколько программных продуктов [38, 39].

Отдельное направление исследований составляет использование онтологий верхнего уровня или общезначимых онтологий (онтологий, не ориентированных на конкретную предметную область) для разработки онтологий в конкретных предметных областях. В качестве такой общей онтологии при разработке предметно-ориентированных онтологий для автоматической обработки текстов часто используется лингвистическая онтология WordNet [11, 12, 35].

Близкие по смыслу понятия общей и предметно-ориентированной лингвистической онтологий могут состоять между собой в следующих отношениях [6, 11, 35]:

– синонимы, то есть понятия двух онтологий, могут быть склеены между собой;

– понятие конкретной онтологии является видом для понятия общей онтологии;

– понятия конкретной онтологии и общей онтологии являются квазисинонимами, то есть одному понятию общей онтологии соответствуют два понятия частной онтологии, или одному понятию частной онтологии соответствуют два понятия общей онтологии. В случае WordNet наличие в нем двух понятий (синсетов), относящихся к одному понятию предметной онтологии, обычно связано с более детальной трактовкой лингвистических явлений, чем это обычно принято в терминологических ресурсах.

В начале работ над Онтологией по естественным наукам и технологиям мы выгрузили часть Тезауруса РуТез – лингвистической онтологии – в предварительную версию новой онтологии. Таким образом, фрагменты общезначимой онтологии были перемещены в другой контекст – область естественных наук. При этом приложение онтологий является одинаковым – информационно-поисковые задачи, такие, как индексация и поиск документов, автоматическая рубрикация, поиск ответов на вопросы, поиск похожего документа и т. п.

В течение почти двух лет эксперты по знаниям работали над Онтологией по естественным наукам и технологиям, дополняли и изменяли полученные понятийные описания. И теперь у нас есть возможность изучить, что изменилось в структуре и отношениях понятий при перемещении их в другие, более специфические предметные области.

Для изучения описаний понятий, перенесенных из Тезауруса РуТез (далее онтология-прототип), мы образовали список таких понятий, которые эксперты одобрили для включения в Онтологию по естественным наукам и технологиям, то есть сняли пометку «понятие-кандидат» (рис. 2). Таких понятий оказалось 4540.

На рис. 2 представлен экран ввода характеристик понятия *АЗУРИТ*. На экране показан список возможных предметных областей, подмножество которых выбирается для характеризации каждого понятия.

С описаниями понятий могли произойти следующие типы изменений:

1) изменение названия понятия;
2) изменение набора текстовых входов понятия:
   а) удаление текстовых входов понятия,
   б) добавление текстовых входов понятия;
3) изменение отношений между понятиями онтологии-прототипа:
   а) исчезновение отношений между понятиями онтологии-прототипа,
   б) появление новых отношений между понятиями онтологии-прототипа;

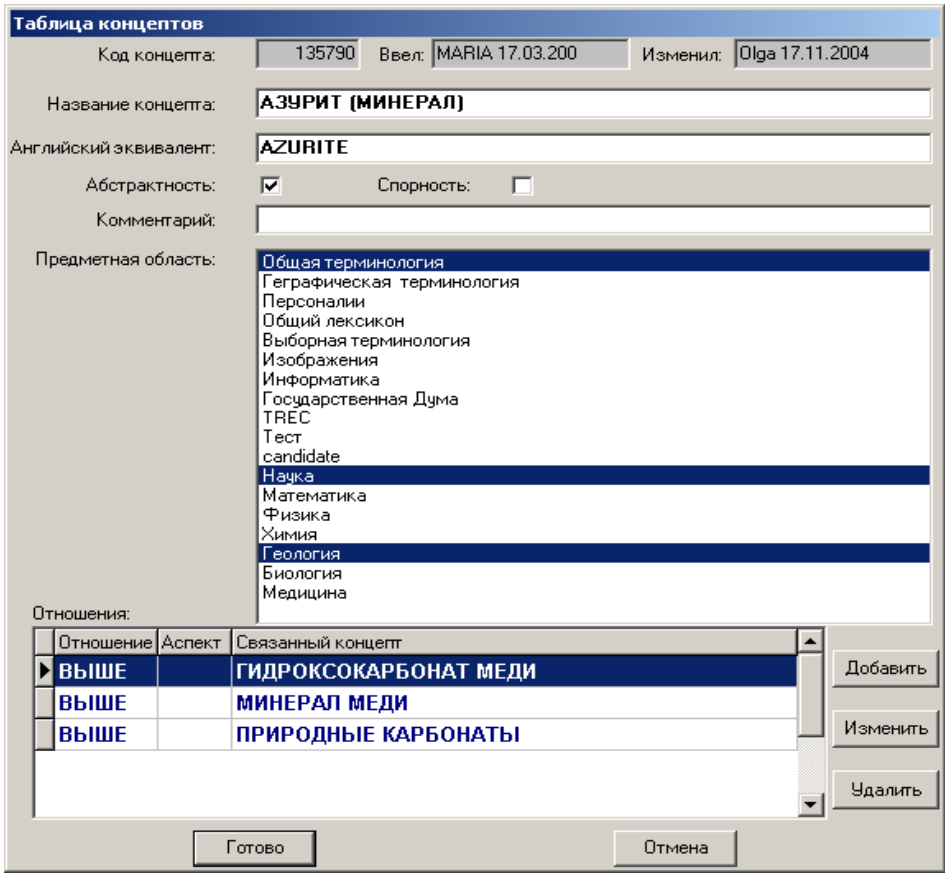

Рис. 2. Экранная форма задания понятия и его ассоциирования с предметными областями

4) введение отношений понятий онтологии-прототипа с новыми понятиями:
   а) введение отношений вверх по иерархии,
   б) введение отношений вниз по иерархии.

В следующих пунктах рассмотрим наиболее интересные явления, которые удалось выявить.

5.1. Удаление текстовых входов понятия. Изменения набора текстовых входов понятия связаны в основном с двумя причинами.

Во-первых, от понятия отсоединяются текстовые входы, носящие метафорический, образный характер, которые свойственны газетным текстам и не употребляются в научной речи, например, *ВЕРБЛЮД – корабль пустыни.*

Во-вторых, (и таких удаленных текстовых входов большинство) часть текстовых входов исходного одного понятия перешла как текстовые входы к новообразованному понятию, то есть понятие практически расщепилось на два (или более) понятий. Например, были разделены в отдельные понятия бывшие синонимы (текстовые входы одного и того же понятия): *ХИМИЧЕСКАЯ РЕАКЦИЯ* и *ХИМИЧЕСКИЙ ПРОЦЕСС, СУДОРОГА* и *СПАЗМ, СОЛИ ФОСФОРНЫХ КИСЛОТ* и *ФОСФАТЫ* и т. п.

5.2. Замена отношений между понятиями онтологии-прототипа на более длинные цепочки отношений. Авторы [6, 35], работавшие с двумя онтологиями, одна из которых более общая, а вторая относится к конкретной предметной области, предполагали, что набор вышестоящих отношений более общей онтологии не подвергается изменениям.

Однако наше сопоставление показало значимое число удаленных родовидовых отношений между понятиями онтологии-прототипа. Более тщательный анализ показал, что достаточно часто удаляемое отношение заменяется на более длинную цепочку отношений, состоящую из двух или трех отношений, то есть между понятиями, перешедшими из более общей онтологии, вклиниваются одно-два понятия из предметной онтологии.

Например, в Тезаурусе РуТез для понятия *АДСОРБЕНТ* было установлено родовидовое отношение к понятию *ВЕЩЕСТВО*, а в новой онтологии создана цепочка понятий *АДСОРБЕНТ – СОРБЕНТ – ВЕЩЕСТВО*.

Отношение между понятиями *БОКСИТ – ГОРНАЯ ПОРОДА* заменилось на цепочку *БОКСИТ – БИОГЕННАЯ ГОРНАЯ ПОРОДА – ОСАДОЧНАЯ ГОРНАЯ ПОРОДА – ГОРНАЯ ПОРОДА*.

Отношение между понятиями *БУЙВОЛ – ЖВАЧНОЕ ЖИВОТНОЕ* заменилось на цепочку *БУЙВОЛ – ПОЛОРОГИЕ – ЖВАЧНОЕ ЖИВОТНОЕ* и т. д.

Количество таких замен одного отношения на цепочку отношений оценивается на текущий момент как более 1000 единиц, что для множества рассматриваемых понятий онтологии-прототипа (4540) представляется значительной величиной.

Важно отметить, что часть из нововведенных отношений может быть перенесена и в исходную онтологию, послужить для уточнения исходных описаний. Вместе с тем значительная часть нововведений не подлежит переносу в онтологию-прототип (см. примеры выше), поскольку введенные понятия соответствуют исключительно научной терминологии и практически не используются в общезначимых текстах.

5.3. Несоответствие наивной, бытовой картины мира и научной картины мира. Тезаурус РуТез предназначен для обработки общезначимых документов: информационных сообщений, нормативных документов, газетных статей. Поэтому он должен отражать знания о мире, которыми обладают авторы и читатели такого вида документов. Картина мира, представленная в тезаурусе, может отличаться от картины мира, излагаемой в рамках естественных наук.

Хрестоматийным примером отличия бытовой картины мира и научной картины мира является знание о том, что кит является млекопитающим, а не рыбой [40]. Однако этому вопросу уделяется достаточное внимание в курсе зоологии средней школы. В частности, не удалось найти ни одного такого текста в текстовой коллекции Университетской информационной системы РОССИЯ (www.cir.ru, более миллиона документов), в котором бы автор считал, что кит – это рыба. Тезаурус РуТез также описывает китов как морских млекопитающих.

Однако удалось выявить ряд несоответствий наивной картины мира, зафиксированной в Тезаурусе РуТез, и научной картиной мира.

Здесь можно выделить два типа различий. Первый тип различий заключается в следующем: то, что в наивной картине мира кажется связанным простым отношением (например, родовидовым), в научной картине мира напрямую не связано. Второй тип различий – то, что представляется несвязанным в наивной картине мира, непосредственно связано между собой в научной картине мира.

Большинство примеров несоответствий находится в сфере биологии. Так, птица эму, которую часто называют *страус эму*, по биологической классификации не является *страусом*.

С другой стороны, по биологической классификации *бледная поганка* относится к *мухоморам*, а *горчица* и *брюква* – к *роду капуста*.

Наиболее запутанной ситуацией является ситуация с употреблением слова *орех*. Биологическая наука рассматривает орех как особый вид плода, к которым, например, не относятся грецкие орехи. Одновременно существует «хозяйственный» (по выражению Большой Советской энциклопедии) взгляд на орехи – плоды деревьев и кустарников, «состоящие из сухой деревянистой оболочки и заключённого в ней съедобного и питательного ядра».

Кроме того, существует еще более отличающееся от научного употребление слова *орех*, которое включает в *орехи арахис, земляной орех*. Это растение по биологической классификации относится к бобовым культурам и не является деревом или кустарником.

Работа с такими несоответствиями связана с двумя видами деятельности: изменение отношений между понятиями на более научно-мотивированные (в том числе и в онтологии-прототипе) и/или ввод разных понятий для разного употребления того или иного слова и описание такого слова как многозначного. Так, видимо, целесообразно иметь два понятия для плода орех – орех как плод ореховых культур (биологическая картина мира) и орех как плод орехоплодных культур («хозяйственная» картина мира).

5.4. Смена антропоцентрической картины мира на естественно-научную картину мира. Наивная картина мира отличается еще и тем, что она ставит в свой центр человека и общество, то есть является антропоцентрической. При переходе к естественно-научной картине мира эта антропоцентричность пропадает, что находит отражение в отношениях онтологии.

Мы заметили это явление в двух проявлениях.

Есть знание, которое известно и в наивной картине мира, но из-за того, что в повседневной жизни некоторая сущность чаще всего встречается в той или иной форме, то эта форма и считается основной для сущности.

Это явление хорошо видно на примере веществ и их агрегатных состояний и проявляется уже в различиях в толкованиях, которые даются в толковых словарях и энциклопедических словарях.

Так, в толковом словаре [41] первое значение слова *вода* таково: *бесцветная прозрачная жидкость, представляющая собою химическое соединение водорода и кислорода и содержащаяся в атмосфере, почве, живых организмах и т. п.*

В Большой Советской энциклопедии термин *вода* имеет такое определение: *окись водорода,* H_2O*, простейшее устойчивое в обычных условиях химическое соединение водорода с кислородом (11.19% водорода и 88.81% кислорода по массе), молекулярная масса 18.0160; бесцветная жидкость без запаха и вкуса (в толстых слоях имеет голубоватый цвет).*

Как следствие, в тезаурусе РуТез установлено отношение *ВОДА – ЖИДКОСТЬ*, в Онтологии по естественным наукам *ВОДА* – это *СОЕДИНЕНИЕ КИСЛОРОДА С ВОДОРОДОМ*, *ОКСИД НЕМЕТАЛЛА*. Вводится дополнительное понятие *ЖИДКАЯ ВОДА* (вода в жидкой фазе, вода в жидком состоянии), которая и является видом понятия *ЖИДКОСТЬ*.

При этом образованным современникам отлично известно, что *СОЕДИНЕНИЕ ВОДА* бывает в разных агрегатных состояниях, но установить отношение между понятиями *ВОДА* и *ЖИДКОСТЬ* в общезначимом ресурсе удобно, так как жидкое агрегатное состояние воды является наиболее обсуждаемым, другие агрегатные состояния *ПАР* и *ЛЕД* воспринимаются как производные от основного.

Еще один элемент антропоцентрической картины мира в тезаурусе РуТез – это наличие таких оценочных понятий, как *СТИХИЙНОЕ БЕДСТВИЕ*, которое оценивает воздействие тех или иных явлений на человеческое существование и включает такие понятия, как *ЗЕМЛЕТРЯСЕНИЕ*, *СМЕРЧ*, *НАВОДНЕНИЕ* и др. Как представляется естественно-научная онтология должна избегать таких оценочных понятий, как *СТИХИЙНОЕ БЕДСТВИЕ*, и должна использовать нейтральные классификации: *СЕЙСМИЧЕСКОЕ ЯВЛЕНИЕ*, *МЕТЕОРОЛОГИЧЕСКОЕ ЯВЛЕНИЕ* и т. п.

5.5. Пример. В качестве примера сравним описание понятия *АЗУРИТ* в составе Тезауруса РуТез и Онтологии по естественным наукам и технологиям.

Азурит – достаточно известный минерал, используется для получения меди и медного купороса, а также для изготовления синей краски.

Описание понятия АЗУРИТ в тезаурусе РуТез таково:

АЗУРИТ
- син АЗУРИТ
- син МЕДНАЯ ЛАЗУРЬ

ВЫШЕ *МИНЕРАЛ*
- син МИНЕРАЛ
- син МИНЕРАЛЬНОЕ ВЕЩЕСТВО
- син МИНЕРАЛЬНЫЙ

АСЦ1 *МЕДЬ*
- син МЕДНЫЙ
- син МЕДНЫЙ КОНЦЕНТРАТ
- син МЕДЬ
- син МЕДЬСОДЕРЖАЩИЙ

а в Онтологии по естественным наукам:

АЗУРИТ (МИНЕРАЛ)
- син АЗУРИТ
- син МЕДНАЯ ЛАЗУРЬ
- син МЕДНАЯ СИНЬ

ВЫШЕ *ГИДРОКСОКАРБОНАТ МЕДИ*
- син (CUOH)2CO3
- син CU2(CO3)(OH)2
- син CU2CO3(OH)2
- син CUCO3-CU(OH)2
- син ВОДНЫЙ КАРБОНАТ МЕДИ
- син ГИДРОКСОКАРБОНАТ МЕДИ

ВЫШЕ *МИНЕРАЛ МЕДИ*
- син МЕДНЫЙ МИНЕРАЛ
- син МИНЕРАЛ МЕДИ
- син ПРИРОДНАЯ МЕДЬ

ВЫШЕ *ПРИРОДНЫЕ КАРБОНАТЫ*
- син КАРБОНАТНЫЙ МИНЕРАЛ
- син МИНЕРАЛ КЛАССА КАРБОНАТОВ
- син ПРИРОДНЫЕ КАРБОНАТЫ

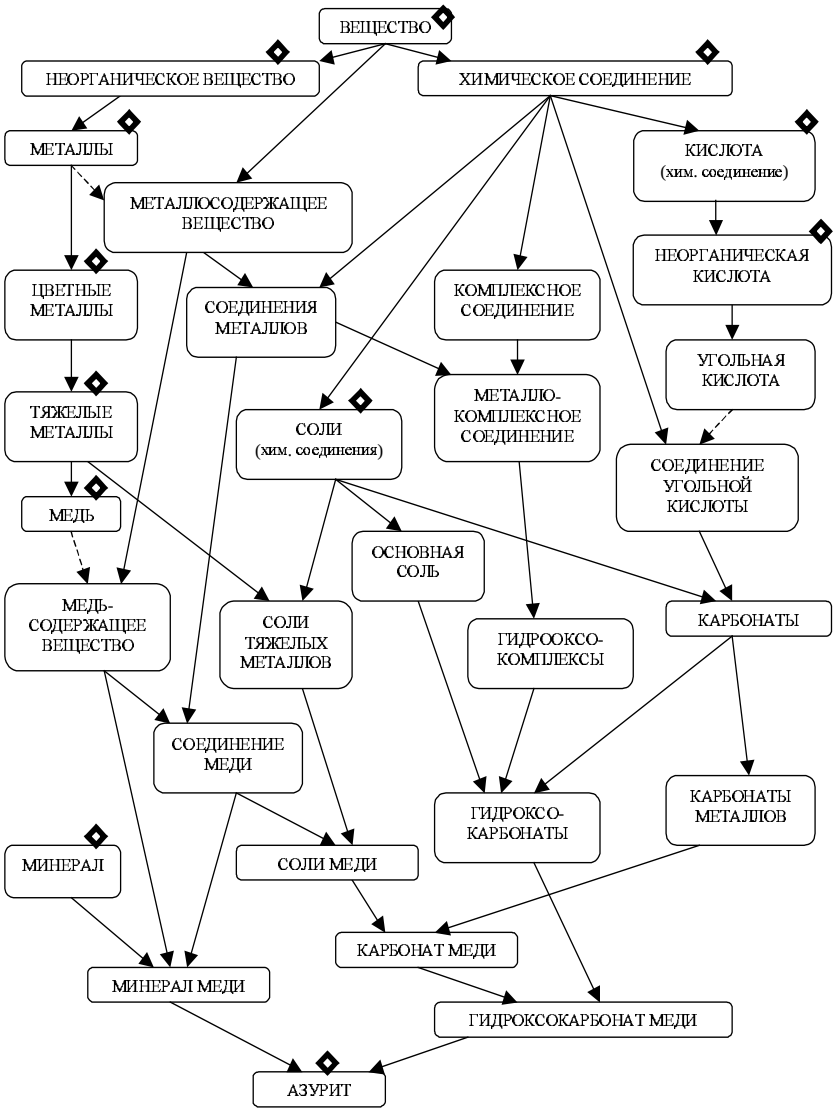

Рис. 3. Фрагмент Онтологии по естественным наукам и технологиям

На рис. 3 показаны верхние уровни иерархии понятия *АЗУРИТ* в Онтологии по естественным наукам и технологиям (за недостатком места не все существующие отношения отражены). Ромбиками помечены понятия, которые были экспортированы из тезауруса РуТез. Мы можем видеть, что прямые отношения понятия АЗУРИТ в тезаурусе РуТез заменились на многоступенчатые структуры, описывающие химический состав минерала.

На рис. 4 для сравнения показаны верхние уровни иерархии понятия *АЗУРИТ* в тезаурусе РуТез.

Заключение

В статье описаны основные принципы и современное состояние разработки Лингвистической онтологии по естественным наукам и технологиям. Разработка онтологии базируется на сочетании подходов к разработке трех разных видов ре-

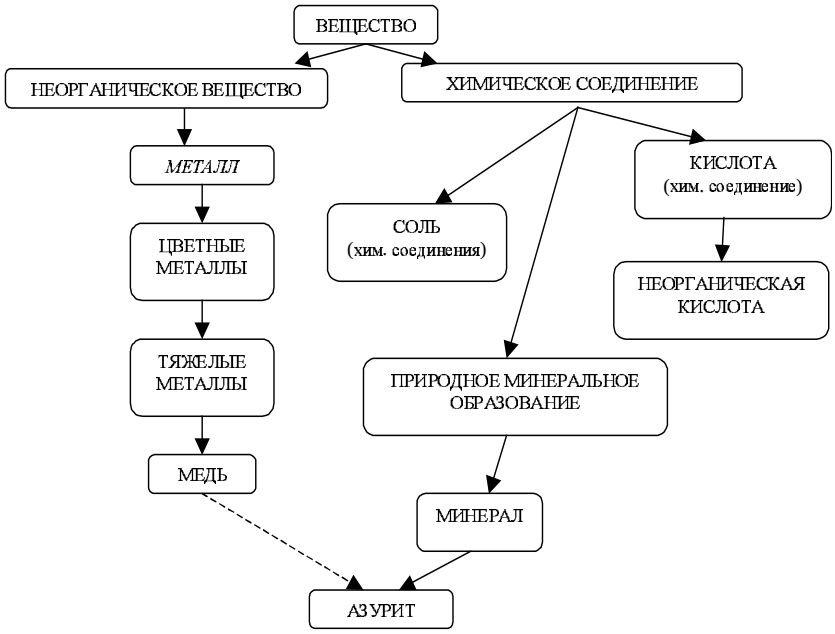

Рис. 4. Аналогичный фрагмент Тезауруса РуТез

сурсов: информационно-поисковых тезаурусов, ресурсов типа WordNet и формальных онтологий.

Сочетание этих трех подходов позволяет нам создавать сверхбольшие лингвистические онтологии для решения задач информационного поиска.

В статье мы подробно рассмотрели типы несоответствий между описаниями одинаковых и близких по смыслу понятий в общезначимой онтологии и предметно-ориентированной онтологии на примерах Тезауруса РуТез как общезначимой онтологии и Онтологии по естественным наукам и технологиям как предметно-ориентированной онтологии.

Одним из наиболее важных выявленных фактов является новый взгляд на структуру «стыка» между такими онтологиями. Стык не представляет собой сплошную полосу понятий, принадлежащих обеим онтологиям. Стык онтологий выглядит как совокупность полос, в которых между уровнями, принадлежащими обеим онтологиям, находятся понятия, принадлежащие только одной из онтологий.

Различия в антропоцентрической «наивной» картине мира и естественно-научной картине мира проявляются в несоответствиях между описаниями понятий в соответствующих онтологиях.

Полагаем, что сложная картина соответствий между описаниями близких по смыслу понятий в онтологии РуТез и Онтологии по естественным наукам и технологиям объясняется тем, что эти две онтологии отличаются по способам рассмотрения внешнего мира. Онтология РуТез рассматривает мир через призму современного цивилизованного общества: что известно о мире значимому количеству образованных людей современного общества, что важно (воздействует, используется) в жизни современного общества. Онтология в области естественных наук и технологий исключает из рассмотрения аспекты общественного мировосприятия и должна описывать в виде онтологической модели устоявшиеся воззрения современной науки, основываясь на материалах научных публикаций.

Summary

B.V. Dobroff, N.V. Loukachevitch. Linguistic ontology on natural sciences and technologies for information-retrieval applications.

The paper describes the main principles of development and current state of Linguistic Ontology on Natural Sciences and Technology intended for information-retrieval problems. In the development of the ontology we combined three different methodologies: development of information-retrieval thesauri, development of wordnets, formal ontology research. Combination of these methodologies allows us to develop large ontologies for broad domains.

Литература

1. *Salton G.* Automatic text processing – the analysis, transformation and retrieval of information by computer. – Addison-Wesley, Reading, MA, 1989.
2. *Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S.* Reengineering thesauri for new applications: the AGROVOC example // J. Digital Information. – 2004. – V. 4, No 4. – Article No 257.
3. *Gruber T.R.* A translation approach to portable ontologies // Knowledge Acquisition. – 1993. – V. 5, No 2. – P. 199–220.
4. *Guarino N.* Formal Ontology and Information Systems // Guarino N. (ed.) Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98. – Trento, Italy: IOS Press, 1998. – P. 3–15.
5. *Guarino N.* Some ontological principles for designing upper level lexical resources // Proc. of First Internat. Conf. on Language Resources and Evaluation (LREC). – Granada, Spain, 1998. – P. 527–534.
6. *Hovy E.H.* Combining and standardizing large-scale, practical ontologies for machine translation and other uses. // Proc. of the First Internat. Conf. on Language Resources and Evaluation (LREC). – Granada, Spain, 1998. – P. 535–542.
7. *Stumme G.* Using ontologies and formal concept analysis for organizing business knowledge // Professionelles Wissensmanagement – Erfahrungen und Visionen. Proc. WM'01 – Shaker, 2001.
8. *Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F.* Ontologies: Expert Systems all over again // AAAI-1999 Invited Panel Presentation. – 1999.
9. *Gomez-Perez A. Fernandez-Lopez M. Corcho O.* OntoWeb. Technical Roadmap. D.1.1.2. – IST project IST-2000-29243. (www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-1-2.pdf).
10. *Miller G., Beckwith R., Fellbaum C., Gross D., Miller K.* Five papers on WordNet. – CSL Report 43. Cognitive Science Laboratory. – Princeton University, 1990.
11. *Buitellar P., Sacalenau B.* Extending Synsets ith Medical Terms. // Proc. of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. – Pittsburg, USA, 2001.
12. *Vossen P.* Extending, Trimming and Fusing WordNet for Technical Documents. // Proc. of WordNet and Other Lexical Resources: Applications, Extensions and Customizations. – Pittsburg, USA, 2001.
13. *Roventini A., Marinelli R.* Extending the Italian WordNet with the Specialized Language of the Maritime Domain. // Proc. of Second International WordNet Conference GWC. – 2004. – P. 193–198.

14. *Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н.* Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Тр. Седьмой Всерос. науч. конф. (RCDL'2005) г. Ярославль, 4–6 окт. 2005 г. – Ярославль: Яросл. гос. ун-т им. П.Г. Демидова, 2005. – С. 70–79.

15. *Добров Б.В., Лукашевич Н.В.* Лингвистическая онтология по естественным наукам и технологиям: основные принципы разработки и текущее состояние // Десятая нац. конф. по искусственному интеллекту с международным участием, Обнинск, 25–28 сент. 2006 г. – М.: Физматлит, 2006. – С. 489–497.

16. *Добров Б.В., Лукашевич Н.В.* Вторичное использование лингвистических онтологий: изменение в структуре концептуализации // Восьмая Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Владимир–Суздаль, 16–18 окт. 2006 г. – С. 56–64.

17. Список нормализованной лексики по экономике и демографии. – М.: АН СССР, ИНИОН, 1989. – Ч. 1. – 169 с.

18. *Шемакин Ю.И.* Тезаурус в автоматизированных системах управления и информации. – М.: Военное изд-во Министерства обороны СССР, 1974. – 192 с.

19. UNBIS Thesaurus, English Edition, Dag Hammarskjold Library of United Nations. – N. Y., 1976.

20. Legislative Indexing Vocabulary. Congressional Research Service. The Library of Congress. Twenty-first Edition, 1994.

21. *Voorhees E.* Natural Language Processing and Information Retrieval // Pazienza M.T. (ed.). Information Extraction: Towards Scalable, Adaptable Systems, – N. Y.: Springer, 1999. – P. 32–48.

22. *Белоногов Г.Г., Зеленков Ю.Г., Кузнецов Б.А., Новоселов А.П., Хорошилов А.А., Хорошилов А.А.* Автоматизация составления и ведения словарей для систем фразеологического перевода с русского языка на английский и с английского на русский // НТИ. Сер. 2. – 1993. – № 12.

23. *Tudhope D., Alani H., Jones Cr.* Augmenting Thesaurus Relationships: Possibilities for Retrieval // J. Digital Libraries. – 2001. – V. 1, No 8.

24. *Gangemi A., Guarino N., Masolo C., Oltramari A.* Understanding Top-Level Ontological Distinctions // Proc. of IJCAI 2001 workshop on Ontologies and Information Sharing. – 2001.

25. *Добров Б.В., Лукашевич Н.В.* Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всерос. конф. по электронным библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Петрозаводск, 2001. – С. 78–82.

26. *Лукашевич Н.В, Добров Б.В.* Отношения в онтологиях для решения задач информационного поиска в больших разнородных текстовых коллекциях // Девятая нац. конф. по искусственному интеллекту с международным участием КИИ2004. Тр. конф.: в 3 т. – М.: Физматлит, 2004. – Т 2. – С. 544–551.

27. *Гак В.Г.* Лексическое значение слова // Лингвистический энциклопедический словарь. – М.: Сов. энцикл., 1990.

28. *Climent S., Rodriguez H., Gonzalo J.* Definitions of the links and subsets for nouns of the EuroWordNet project. – Deliverable D005, WP3.1, EuroWordNet, LE2-4003, 1996.

29. *Mahesh K., Nirenburg S.* A Situated Ontology for Practical NLP. // Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95). – Montreal, Canada, 1995.

30. *Гаврилова Т.А.* Извлечение знаний: лингвистический аспект // Корпоративные системы. – 2001. – Т. 10, № 25. – С. 24–28.

31. *Лукашевич Н.В.* Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер. 2. – 1995. – № 3. – С. 21–24.

32. *Лукашевич Н.В., Салий А.Д.* Тезаурус для автоматического рубрицирования и индексирования: разработка, структура, ведение // НТИ. Сер. 2. – 1996. – № 1. – С. 1–6.

33. *Добров Б.В., Лукашевич Н.В.* Построение и использование тематического представления содержания документов // V нац. конф. с международным участием «Искусственный интеллект-96». – Казань, 1996. – Т. 1. – С. 130–134.

34. *Добров Б.В., Лукашевич Н.В., Сыромятников С.В.* Формирование базы терминологических словосочетаний по текстам предметной области // Пятая Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», СПб., 28–31 окт. 2003 г. – СПб.: С.-Петерб. ун-т, 2003. – С. 201–210.

35. *Magnini B., Speranza M.* Merging Global and Specialized Linguistic Ontologies // Proc. of OntoLex 2002.

36. *Guarino N.* Understanding, building and using ontologies. Int. // J. Human-Computer Studies (IJHCS). – 1997. – V. 46. – P. 293–310.

37. *Kalinichenko L., Skvortsov N.* Ontology reconciliation in terms of type refinement // Proc. of the 6th Russian Conference on Digital Libraries RCDL2004, Pushchino, Russia, September 2004.

38. *McGuinness D.L., Fikes R., Rice J., Wilder S.* An environment for merging and testing large ontologies // Proc. of the Seventh International Conference (KR'2000). Morgan Kaufmann Publishers, San Francisco.

39. *Noy N.F., Musen M.A.* PROMPT: Algorithm and tool for automated Ontology merging and alignment // Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin, TX.

40. *Апресян Ю.Д.* Лексическая семантика. Синонимические средства языка. – М.: Языки рус. культуры, Изд. фирма «Вост. лит.» РАН, 1995. – 472 с.

41. *Ефремова Т.Ф.* Современный толковый словарь русского языка: в 3 т. – М.: АСТ, 2006.

Поступила в редакцию
04.09.07

Добров Борис Викторович – кандидат физико-математических наук, заведующий лабораторией Научно-исследовательского вычислительного центра Московского государственного университета им. М.В. Ломоносова.

E-mail: *dobroff@mail.cir.ru*

Лукашевич Наталья Валентиновна – кандидат физико-математических наук, старший научный сотрудник Научно-исследовательского вычислительного центра Московского государственного университета им. М.В. Ломоносова.

E-mail: *louk@mail.cir.ru*