Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска[♣]

© Б.В. Добров
Научно-исследовательский вычислительный центр МГУ;
АНО Центр информационных исследований
dobroff@mail.cir.ru

Н.В. Лукашевич
Научно-исследовательский вычислительный центр МГУ;
АНО Центр информационных исследований
louk@mail.cir.ru

М.Н. Синицын
Научно-исследовательский вычислительный центр МГУ;
ГУ НИМЦ "Базис" Минобразования России

В.Н. Шапкин
ГУ НИМЦ "Базис" Минобразования России

Аннотация

В работе 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 14,400 «подтвержденных» экспертами понятий с установленными онтологическими связями, 34,000 текстовых входов для таких научных дисциплин как математика, физика, химия, геология и биология.

1 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.

Труды 7[ой] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

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 [29, 30].

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

Согласно, [22] под онтологиями понимают систему явной концептуализации предметной области, то есть формального представления предметной области.

Отметим, что существуют разные формальные интерпретации [19, 24, 31] 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 [22].

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 [36]:

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 [21] 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.) [21].

В качестве примера лингвистической онтологии часто приводится ресурс WordNet [27]. Этот ресурс представляет в виде иерархической структуры систему значений слов общезначимого английского языка. Вместе с тем возникает достаточно много проектов, которые описывают на основе модели WordNet терминологические системы конкретных предметных областей, то есть создают лингвистические онтологии этих областей [16, 28, 35].

Коллектив авторов в 2004 году начал работу над созданием лингвистической онтологии для автоматической обработки в области естественных наук.

В статье описывается идея, методология и текущее состояние проекта. Во втором разделе рассматриваются достоинства и недостатки существующих лингвистических ресурсов с точки зрения применимости для автоматической обработки текстов. В третьем разделе мы описываем идею и основные положения проекта создания лингвистической онтологии для естественных наук. В следующем разделе описывается ранее созданная авторами онтология - Общественно-политический тезаурус, на основе методологии создания которой создается и новая лингвистическая онтология. В разделах 5 и 6 излагаются этапы построения онтологии, приводятся данные о текущем состоянии ресурса.

2 Ресурсы для смыслового анализа электронных коллекций

2.1 Традиционные информационно-поисковые тезаурусы

Хронологически первыми ресурсами, служащими для упорядочения работы с электронными коллекциями были информационно-поисковые тезаурусы (ИПТ) [12, 15, 25, 33], 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 объем потоков информации в настоящее время 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 [29, 34].

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.

Чтобы использоваться в автоматическом режиме традиционным тезаурусам не достает 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 свойств, базирующая в основном на использовании отношений ВЫШЕ-НИЖЕ и Ассоциация.

В России наиболее известен Тезаурус научно-технических терминов [15], который издан в 1972 году. Тезаурус описывает терминологию военно-промышленного комплекса 70-х годов, не соответствует реалиям и технологиям настоящего времени. ВИНИТИ обладает громадным массивом научно-технических текстов, имеются наборы терминов [1] по научно-техническим отраслям. Но эти термины не организованы иерархическими связями в единый ресурс научно-технической терминологии.

2.2 От информационно-поисковых тезаурусов к фундаментальным онтологиям

Некоторые авторы [30, 32], 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 [30] в качестве примеров модификации информационно-поискового тезауруса по сельскому хозяйству AGROVOC приводятся следующие словарные статьи:

Исходные статьи тезауруса (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 отработал, помимо изменений в описании понятий и терминов предметной области, нужно 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 этого 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 надежно можно было 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.

2.3 Отношения в онтологии, применяемой в неопределенных контекстах

На основе анализа, проведенного в предыдущей секции, можно заметить, что информационно-поисковые онтологии в течение долгого времени будут вынуждены 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, т.е. 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 невозможности использования сложных правил вывода, для осуществления вывода по тексту желательно найти другие типы отношений, обладающие свойствами транзитивности и наследования, подобно таксономическим отношениям.

Как представляется, что именно такого рода отношениями являются отношения онтологической зависимости, изучаемые в рамках философской дисциплины «формальная онтология» [20].

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

- подразумевает ли существование сущности существование чего-либо еще (строгая зависимость – rigid dependence), например, *кипение* не возможно без существования конкретного объема жидкости, которая кипит;
- предполагается ли существование примеров некоторого класса (родовая зависимость - generic dependence) некоторых сущностей, как 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, существования 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 эти работы заканчиваются, а солома длительное время продолжает существовать.

В работе [18] постулируется транзитивность отношений онтологической зависимости.

В работах [6, 10] было показано, что отношения строгой и родовой онтологической зависимости эффективны для создания ресурсов для информационного поиска.

2.4 WordNet как лингвистическая онтология

Целью разработки WordNet [27] не являлось 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 уже понятия в том смысле, что включает лишь различительные черты объектов, а понятия охватывают их наиболее глубокие существенные свойства…» [3].

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 отношение *антоним, дериват, валентности* [17], 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, 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 [26].

3 Проект разработки новой лингвистической онтологии

В проекте предлагается создать лингвистическую онтологию 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.

Из 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.

Из методологии разработки лексических ресурсов типа WordNet важны следующие положения:

- многоступенчатое иерархическое построение лексико-терминологической системы понятий;
- технология описания значений многозначных слов и выражений.

Из методологии разработки формальных онтологий:

- разработка лингвистической онтологии как иерархической системы понятий;
- строгость построения таксономии, отличие истинно таксономических отношений от ролевых отношений;
- использование для описания нетаксономических отношений онтологической зависимости.
- в качестве аксиом (правил вывода) использовать свойства транзитивности и наследования таксономических отношений и отношений онтологической зависимости.

Основной процедурой разработки такой лингвистической онтологии является следующая совокупность 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 [4], 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 доклада в разработке онтологии принимают участие эксперты-лингвисты: Штернова О.А., Селиванова Т.М, Каргина И.А.

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.

Действительно, подробные сетки понятий, описываемые с единых всем понятных “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.

4 Отправная точка

Авторы проекта ранее [8, 9] создали информационно-поисковый тезаурус для автоматического индексирования текстов в общественно-политической области (далее – Общественно-политический тезаурус), включающих более 32 тысяч понятий, 79 тысяч русскоязычных и 80 тысяч 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 [5, 6] 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.

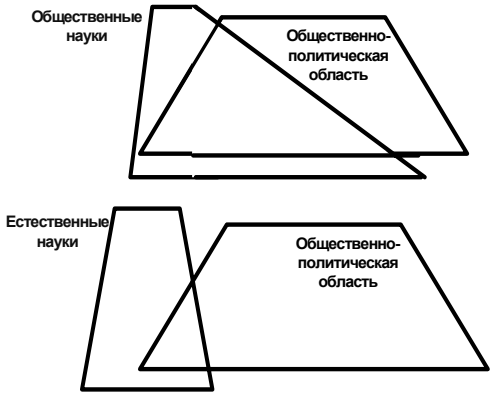

Рис.1. Научная лексика и терминология в общественно-политическом тезаурусе

Для реализации обсуждаемого проекта наиболее важны созданные ранее технологии быстрого автоматизированного формирования [7] 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 (Рис.1). Тем не менее сфера естественных наук затрагивается в связи с обсуждением вопросов промышленности, нефтедобычи, медицины и т.п., поэтому соответствующая научная лексика и терминология неплохо представлена в тезаурусе, что позволило поставить задачу ее использования при создании нового ресурса.

5 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.

5.1 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 [7].

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 Общественно-политического тезауруса могли выступать «зародышами» для формирования более длинных словосочетаний.

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

5.2 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.

5.2.1. Отбор новой терминологии

По каждой предметной области были образованы верхние части частотных списков терминоподобных слов (по 10 тысяч) и словосочетаний (по 15 тысяч), 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 тысяч помеченных слов и словосочетаний.

5.2.2. 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 были заданы несколько понятий верхнего уровня, такие как *=НАУКА=*, *=РАСТЕНИЕ=* и т.п., касающиеся 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 помечены специальными пометками отнесения к дополнительной предметной области соответствующей науки и к специальной служебной рабочей предметной области «кандидат».

5.2.3. 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 не известен, 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.

5.2.4. Замыкание предметной области

Для отобранных из Общественно-политического тезауруса понятий (получивших пометку «кандидат») было выполнено «замыкание» - были добавлены понятия, расположенные выше по таксономическим связям. Эти понятия получали аналогичные дополнительные пометки предметных областей.

5.2.5. 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

Рис.2. Визуализация результатов обработки текста «Функции клетки»

отнесения к служебной предметной области «кандидат», то мы использовали стандартную процедуру «экспорта» фрагмента тезауруса для формирования нового ресурса.

Размер первой версии ресурса составил более 70 тысяч понятий.

После чего данные версии были выданы экспертам, ответственным за формирование фрагмента лингвистической онтологии по отдельным научным дисциплинам.

5.3 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.5.2.3).

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.

В настоящее время экспертами используется следующих три основных источника:

- профильные и общие энциклопедии [2, 11, 13, 14], 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 Интернет. Такая проверка производится с использованием глобальных поисковых машин.

6 Текущее состояние проекта

Численные характеристики развития проекта приведены в Таблицах 1 и 2 (здесь ОПТ – обозначает Общественно-политический тезаурус без учета географической лексики и терминологии).

Географические понятия и лексика - 6,7 тысяч понятий, 10 тысяч текстовых входов – добавляются в получающийся ресурс из ОПТ.

Как можно видеть из представленных данных, в результате труда экспертов число «кандидатов» уменьшается, одновременно растет размер ресурса.

Отметим, что в настоящее время около 30% нового ресурса получено из ОПТ.

Таблица 1. Покрытие понятийной структуры предметной области

	2004, окт. тыс.	2005, июнь тыс
Всего понятий	62,7	65,0
из них из ОПТ	24,3	24,3
«Кандидаты»	56,1	43,7
Науки (без «канд.»)	**--**	**14,4**
из них из ОПТ	--	4,0
Итого, вкл. географ.	**6,6**	**21,1**

Таблица 2. Покрытие терминологии предметной области

	2004, окт. тыс.	2005, июнь тыс
Всего терминов	116,7	132,7
из них из ОПТ	74,0	74,0
Кандидаты»	106,8	88,2
Науки (без «канд.»)	**--**	**34,2**
из них из ОПТ	--	12,0
Итого, вкл. географ.	**9,7**	**44,5**

В своем современном состоянии полученный ресурс может быть использован в ранее созданном программном обеспечении обработки текстов [5, 6] для определения тематики обрабатываемых естественнонаучных текстов. На Рис.2 представлен пример результатов обработки, в частности построенная по тексту тематическая аннотация. Отмечены найденные текстовые входы, а также понятия, связанные с понятием =*ТКАНЬ*=.

Заключение

Как представляется авторам, реализация проекта позволит достичь следующих результатов:

1) будет разработана лингвистическая онтология большого размера для автоматической обработки текстов научно-технической тематики;
2) ресурс будет бесплатен для некоммерческого применения.

Развитие ресурса будет обеспечиваться продажей лицензий для коммерческого применения, а также выполнением хоздоговорных работ для создания специальных информационных систем, особо точным и полным покрытием терминологии в специфических областях и т.п.

Литература

[1] Белоногов Г.Г., Зеленков Ю.Г., Кузнецов Б.А., Новоселов А.П., Хорошилов Ал-др А., Хорошилов Ал-сей 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.

[2] Биология: Энциклопедия / Под ред. М.С.Гилярова. – М: Большая Российская энциклопедия, 2003. – 864 с.

[3] Гак В.Г., Лексическое значение слова – Лингвистический энциклопедический словарь. – М: Советская энциклопедия. – 1990.

[4] Гаврилова Т.А., Извлечение знаний: лингвистический аспект //Корпоративные системы. - 2001.- N10 (25), с.24-28.

[5] Добров Б.В., Лукашевич Н.В., Построение и использование тематического представления содержания документов // V национальная конференция с международным участием "Искусственный интеллект-96", Казань, 1996, Том I, С.130-134.

[6] Добров Б.В., Лукашевич Н.В., Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по Электронным Библиотекам "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - Петрозаводск, 2001 - С.78-82.

[7] Добров Б.В., Лукашевич Н.В., Сыромятников С.В., Формирование базы терминологических словосочетаний по текстам предметной области // Пятая Всероссийская научная конференция "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Санкт-Петербург, 28 -31 октября 2003 г. - СПб.: СпбГУ - 2003. - С.201-210.

[8] Лукашевич Н.В., Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. - 1995. - N 3. - С.21-24.

[9] Лукашевич Н.В., Салий А.Д., Тезаурус для автоматического рубрицирования и индексирования: разработка, структура, ведение // НТИ. Сер.2. - 1996. - N 1. - С.1-6. 3.

[10] Лукашевич Н.В, Добров Б.В., Отношения в онтологиях для решения задач информационного поиска в больших разнородных текстовых коллекциях // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции. В 3-х т. - Т2. – М.: Физматлит, 2004. – С.544-551.

[11] Математика: Энциклопедия / Под ред. Ю.В.Прохорова. – М: Большая Российская энциклопедия, 2003. – 845 с.

[12] Список нормализованной лексики по экономике и демографии. - М.: АН СССР, ИНИОН, 1989.- Ч. 1. - 169 с.

[13] Физика: Энциклопедия / Под ред. Ю.В.Прохорова. – М: Большая Российская энциклопедия, 2003. – 944 с.

[14] Химия: Энциклопедия / Под ред. И. Л. Кнунянца – М: Большая Российская энциклопедия, 2003. – 972 с.

[15] Шемакин Ю.И., Тезаурус в автоматизированных системах управления и информации. - М: Военное изд-во министерства обороны СССР, 1974. - 192 с.
[16] Buitellar, P., Sacalenau, B., Extending Synsets ith Medical Terms. // Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA, 2001.
[17] 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.
[18] Gangemi A., Guarino N., Masolo C., Oltramari A., Understanding Top-Level Ontological Distinctions // Proceedings of IJCAI 2001 workshop on Ontologies and Information Sharing, 2001.
[19] Guarino,N. Formal Ontology and Information Systems. In N. Guarino, editor, Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, Trento, Italy, pages 3-- 15. IOS Press, June 1998.
[20] Guarino N., Some Ontological Principles for Designing Upper Level Lexical Resources. // Proceedings of First International Conference on Language Resources and Evaluation, 1998.
[21] 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)
[22] Gruber T.R., A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993.
[23] Hirst G. Ontology and the Lexicon. - Handbook on Ontologies in Information Systems, Berlin – Springer, 2003.
[24] Hovy E.H., Combining and standardizing large-scale, practical ontologies for machine translation and other uses. // Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). Granada, Spain, 1998.
[25] Legislative Indexing Vocabulary. Congressional Research Service. The Library of Congress. Twenty-first Edition, 1994.
[26] 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), 1995. Montreal, Canada.
[27] Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K., Five papers on WordNet. - CSL Report 43. Cognitive Science Laboratory, Princeton University, 1990.
[28] Roventini A., Marinelli R. Extending the Italian WordNet with the Specialized Language of the Maritime Domain. // Proceedings of Second International WordNet Conference GWC – 2004. – pp. 193-198.
[29] Salton G., Automatic Text Processing - The Analysis, Transformation and Retrieval of Information by Computer. Addison-Wesley, Reading, MA, 1989.
[30] Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S. Reengineering Thesauri for New Applications: the AGROVOC Example. - Article No. 257, 2004-03-17.
[31] [Stumme,2001] Stumme G. Using ontologies and formal concept analysis for organizing business knowledge. // Proc. Referenzmodellierung 2001.
[32] Tudhope D., Alani H., Jones Cr. Augmenting Thesaurus Relationships: Possibilities for Retrieval. – Journal of Digital Libraries. Volume 1, Issue 8. – 2001
[33] UNBIS Thesaurus, English Edition, Dag Hammarskjold Library of United Nations, New York, 1976.
[34] Vorhees E., Natural Language Processing and Information Retrieval.
[35] Vossen, P.: Extending, Trimming and Fusing WordNet for Technical Documents. // Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA, 2001.
[36] Welty, C., McGuinness, D., Uschold, M., Gruninger, M., and Lehmann, F. Ontologies: Expert Systems all over again. AAAI-1999 Invited Panel Presentation. 1999.

Development of Linguistic Ontology on Natural Sciences for Information Retrieval Purposes

Boris V. Dobrov, Natalia V.Loukachevitch, Mikhail N. Sinitsyn, Vladimir N. Shapkin

In the paper we describe an idea, a methodology and the current state of the project on development of a linguistic ontology – a new type of an information-retrieval thesaurus on natural sciences (mathematics, physics, geology, chemistry, biology) intended for automatic conceptual indexing of documents and other information-retrieval tasks. In construction of the ontology we combined three different methodologies: (1) the methods of construction of information-retrieval thesauri (information-retrieval context, analysis of terminology, terminology-based concepts, a small set of relation types); (2) the development of wordnets for various languages (language-motivated concepts, description of ambiguous terms); (3) ontology and formal ontology research (concept-based structure, strictness of relations description, necessity of many-step inference).

♣ Работа частично выполняется при финансовой поддержке РФФИ, грант № 03-01-00472.