ОНТОЛОГИИ КАК СИСТЕМЫ ХРАНЕНИЯ ЗНАНИЙ

Н.С. Константинова, О.А. Митрофанова

Санкт-Петербургский государственный университет,
Факультет филологии и искусств, Кафедра математической лингвистики
199034, г. Санкт-Петербург, Университетская наб., д. 11

Аннотация. Обзор представляет исследовательские результаты, достигнутые в новой области науки, связанной с построением и применением онтологий. В рамках работы освещены различные точки зрения на понятие онтологии, используемого в современных информационных технологиях, дано определение этого термина, а также рассмотрены различные классификации онтологий. Приводится общая характеристика автоматических методов построения онтологий, в том числе методов автоматического выявления аксиом и слияния различных онтологий. Описаны основные языки представления онтологий и наиболее значимые существующие онтологические ресурсы. В обзоре сделана попытка представить в общих чертах методологию построения онтологий, рассмотреть проблемы, сопровождающие создание онтологий, и их возможные решения. Также в данном обзоре упоминаются возможные области применения онтологий в информационных системах.

Annotation. The review presents research results achieved in a new field of knowledge dealing with the development and application of ontologies. The paper describes different approaches to the notion of ontology and discusses various classifications of ontologies. The review gives a general description of automatic techniques of ontology development, in particular, of automatic extraction of axioms and of ontology merging. The principal ontology languages and the most significant contemporary ontological resources. The paper presents an attempt to give a general outline of ontology development techniques, to discuss the problems of ontology

development and their possible solutions. The review also deals with the possible applications of ontologies in informational systems.

Введение

Развитие наукоемких областей человеческой деятельности в современном обществе сопровождается возрастанием роли компьютерных технологий. Сейчас значительно увеличивается поток информации, появилась необходимость поиска новых способов ее хранения, представления, формализации и систематизации, а также автоматической обработки. Таким образом, растет интерес к всеобъемлющим базам знаний, которые возможно использовать для различных практических целей. Огромный интерес вызывают системы, способные без участия человека извлечь какие-либо сведения из текста. Как результат, на фоне вновь возникающих потребностей развиваются новые технологии, призванные решить заявленные проблемы. Наряду с World Wide Web появляется его расширение, Semantic Web, в котором гипертекстовые страницы снабжаются дополнительной разметкой, несущей сведения о семантике включаемых в страницы элементов. Неотъемлемым компонентом Semantic Web является понятие онтологии, описывающее смысл семантической разметки.

В общих чертах под онтологией понимается система понятий некоторой предметной области, которая представляется как набор сущностей, соединенных различными отношениями (подробнее см. раздел 1.1). Онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную область знаний. Преимуществом онтологий в качестве способа представления знаний является их формальная структура, которая упрощает их компьютерную обработку.

Можно говорить о неявном применении онтологий в качестве систем понятий в естественных науках (биология, медицина, геология и другие), где они служат своего рода фундаментом для построения теорий. Поскольку классификационная структура (таксономия) является неотъемлемой частью любой онтологии, можно говорить о присутствии элементов онтологий в специальных классификациях и системах индексации (например, в библиотечных классификационных кодах).

В явном виде онтологии используются как источники данных для многих компьютерных приложений (для информационного поиска, анализа текстов, извлечения знаний и в других информационных технологиях), позволяя более

эффективно обрабатывать сложную и разнообразную информацию. Этот способ представления знаний позволяет приложениям распознавать те семантические отличия, которые являются само собой разумеющимися для людей, но не известны компьютеру.

Само понятие онтологии известно давно, но, будучи переосмысленным, оно стало применяться в компьютерных технологиях лишь недавно. Полноценная разработка онтологий в новом смысле этого термина началось лишь в конце 90-х. Это достаточно новая и мало разработанная отрасль прикладной лингвистики. Большинство работ по созданию и использованию онтологий проводится за рубежом, однако и в России существует ряд исследователей, работающих в этой области, см. также работы по онтологиям, опубликованные в России ([1], [3], [4], [6] и др.). Уже на данном этапе создан ряд обширных онтологий, включающих несколько тысяч понятий: OMEGA, SUMO, DOLCE и другие.

Онтологии широко используются во всех областях, занимающихся обработкой данных на естественном языке. В связи с использованием онтологий в различных приложениях возникла необходимость создания стандартизированных способов их представления. Началось развитие разнообразных языков, которые могли бы применяться повсеместно во всех системах, самыми известными являются RDF и OWL. Возникло также большое количество редакторов для создания, пополнения и изменения онтологий. Каждое из этих средств обычно направлено на работу с определенным форматом данных и обладает своими особенностями.

1.1. Определение понятия «онтология»

Термин «онтология» используется в нескольких областях знания и имеет два различных значения [3]:

- «Философская дисциплина, которая изучает наиболее общие характеристики бытия и сущностей»
- «Онтология – артефакт, структура, описывающая значения элементов некоторой системы»

Термин «онтология» имеет долгую историю в философии, где он еще со времен Аристотеля обозначал науку о бытии. Предметом онтологии выступало само по себе сущее, в рамках этого раздела философии выделялись базовые категории и общие свойства, типы сущностей. Как упоминают С. Ниренбург и В. Раскин [26], в качестве синонима термина «онтология» в его философском понимании часто используется термин «метафизика».

Понятие онтологии в инженерной области знания рождает большое количество дискуссий, в которых каждый автор стремится предложить свое определение. Отчасти это можно объяснить новизной области исследования, с другой стороны, – разнообразием практических задач, решаемых с использованием онтологий.

В основу нашего обзора определений легла работа Н.Гуарино [18], где определения онтологии рассматриваются с разных точек зрения. В современных информационных технологиях наиболее часто упоминается и используется определение онтологии, сформулированное Н.Грубером: «Онтология – это спецификация концептуализации» [17]. Эта дефиниция является своеобразным обобщением, формальной интерпретацией многих других определений. Центральным в нем является понятие «концептуализация», которое было введено в работе [16]. Основная сложность заключаются в том, что термин «концептуализация» имеет разнообразные контексты употребления, поэтому данный термин вызывает разногласия.

Как поясняет Н.Гуарино, этот термин в общих чертах интуитивно понятен, и его четкая формулировка не дается при обсуждении понятия «онтология». Стоит пояснить, что под «концептуализацией» понимается строгое описание системы понятий, объектов

и других сущностей и отношений, связывающих их друг с другом. Можно сказать, что концептуализация – это абстрактное, упрощенное видение мира, который мы хотим представить для каких-то целей. Таким образом, концептуализация расчленяет какую-либо область знаний, существующую в целостном виде, выделяет из этой области отдельные объекты, а затем формулирует отношения, свойственные для данной области. Основная часть формально представленного знания базируется на концептуализации, каждая база знаний или система, основанная на знаниях, явно или неявно связывается с какой-то концептуализацией.

Однако существует две различные трактовки природы концептуализации, их можно охарактеризовать, как интенсиональную и экстенсиональную. Экстенсиональная трактовка, которой придерживаются авторы работы [16], подразумевает, что каждое понятие и отношение может исчерпывающе описываться перечислением индивидуальных сущностей, к которым оно применимо. Н.Гуарино же считает эту точку зрения узкой и обобщает ее, развивая интенсиональный подход. Он предлагает идентифицировать понятия не через их перечисление, то есть экстенсионал, а через их внутренние свойства и характеристики, так называемое «предполагаемое содержание». Этот универсальный подход дает специалисту возможность подводить под одну и ту же концептуализацию разные положения вещей. Таким образом, концептуализация становится относительно независимой от индивидуальных сущностей, необходимым является лишь сохранение заданных типов отношений. Важно внутреннее содержание понятий, а не перечисление соответствующих конкретных индивидов.

Однако приведенное ранее определение, связанное с понятием «концептуализации», является далеко не единственным. В литературе можно также найти определение онтологии как «теории того, какие сущности могут существовать в уме хорошо осведомленного (knowledgeable) агента» [35]. Данное определение выявляет другой подход к этому понятию. Эта формулировка позволяет включать в онтологию набор понятий, но не дает возможность задавать их структуру. Этот пробел в определении особенно значим, так как в рамках искусственного интеллекта в качестве синонима онтологии часто используется понятие «терминология», а в ней структура безусловно содержится.

Объединением выше упомянутых дефиниций становится определение из работы [34], вводящее «онтологию», как «спецификация концептуализации на уровне эксплицитных знаний, зависящее от предметной области или задачи, для которой она предназначена». Таким образом, онтология зависит от определенной точки зрения, однако, как упоминает Н.Гуарино, как раз степень этой зависимости является определяющим фактором для возможности ее многократного использования. А ведь именно в возможности многократного использования онтологий видится их значимость и ценность.

В других определениях, приводимых Н.Гуарино, делается упор на иной аспект онтологий, и они определяются как соглашения о совместно используемых концептуализациях. При этом поясняется, что эти совместные концептуализации включают в себя понятийные структуры для моделирования знания какой-то предметной области. Это своего рода соглашение, какие схемы и теории использовать при описании предметной области. Таким образом, здесь проводится разграничение понятий «онтология» и «концептуализация». Об онтологии говорится уже не как о спецификации концептуализации, а лишь как о соглашении о концептуализации. Однако степень детализации этого соглашения будет напрямую зависеть от предназначения конкретной онтологии и целей, поставленных перед исследователем.

Н. Гуарино упоминает еще один подход к определению «онтологии»: «Онтология – это конкретный артефакт, созданный для выражения значений, подразумеваемых у совместно используемой лексики». Здесь упоминается, что онтология предоставляет средства для передачи подразумеваемого значения.

Н.Гуарино предлагает также определять онтологию, как «логическую теорию, которая ограничивает подразумеваемые модели логического языка». Таким образом, под онтологией понимается нечто большее, чем просто детализированный набор понятий и отношений. В онтологию включаются и ограничения, накладываемые на отношения в рамках данной области. Это некоторый набор аксиом, который строится на базе понятий и отношений между ними. Таким образом, например, в рамках искусственного интеллекта можно описать онтологию программы, определив множество объектов и связав их с описаниями, а также введя формальные аксиомы, которые ограничивают интерпретацию и совместное употребление этих терминов.

Формально онтологию можно назвать формулировкой логической теории, некого исчисления со своими правилами. Эта теория позволяет систематизировать категории действительности и/или выражаемые в языке значения. Следовательно, такое определение онтологии можно считать более широким взглядом на данное понятие, нежели предыдущие.

В качестве рабочего определения, наиболее приспособленного для целей компьютерной лингвистики, можно взять дефиницию, предложенную Эдвардом Хови [20]: «Онтология – это структура данных с заданными в ней символами, позволяющими представлять концептуализации для обработки компьютерными программами».

Рассмотрев оба понимая значения термина «онтология», вводимые в философии и в инженерной области знания, можно обнаружить такое соотношение двух значений, как «процесс-результат». В философии «онтология» - это наука, изучающая бытие, а «онтология» в инженерии – это отображение бытия в формализованном виде. Однако С. Ниренбург и В. Раскин [26] указывают на то, что предмет описания формальной онтологии соотносится внутри философии скорее не с наукой о бытии (онтологией, или метафизикой), а с наукой о познании («гносеологией»). Для формальной онтологии важны не столько сами понятия, сколько использование их людьми, знания людей о данных понятиях. В связи с этим не встает, например, вопрос, правомерно ли включать в формальные онтологии несуществующие сущности (например, вымышленные существа). Метафизика будет утверждать, что в бытии такого не существует, однако, обратившись к гносеологии, мы обнаружим, что это существует в умах людей и на данном основании может включаться в онтологию.

Таким образом, следует заключить, что четкой взаимообусловленности между двумя значениями термина «онтология» – в философии и в инженерии знаний – не прослеживается. Связь между ними носит скорее произвольный ассоциативный характер и не будет обсуждаться далее в нашем исследовании.

1.2. Структура онтологии

Рассмотрев возможные содержательные интерпретации понятия «онтология», остановимся подробнее на структуре онтологии, ее составляющих. В общем виде структура онтологии представляет собой набор элементов четырех категорий:

- понятия;
- отношения;
- аксиомы;
- отдельные экземпляры;

Понятия рассматриваются как концептуализации класса всех представителей некой сущности или явления (например, Животное, Чувство). Классы (или понятия) являются общими категориями, которые могут быть упорядочены иерархически. Каждый класс описывает группу индивидуальных сущностей, которые объединены на основании наличия общих свойств.

Понятия могут быть связаны различного рода отношениями (например, Длина, Местоположение), которые связывают воедино классы и описывают их. Самым распространенным типом отношений, использующимся во всех онтологиях, является отношение категоризации, то есть отнесение к определенной категории. Этот тип отношений имеет ряд других названий [3], встречающийся в различных исследованиях:

- таксономическое отношение;
- отношение IS-A;
- класс – подкласс;
- лингвистика: гипоним – гипероним;
- родовидовое отношение;
- отношение a-kind-of.

Аксиомы задают условия соотнесения категорий и отношений, они выражают очевидные утверждения, связывающие понятия и отношения. Под аксиомой можно понимать утверждение, вводимое в онтологию в готовом виде, из которого могут быть выведены другие утверждения. Они позволяют выразить ту информацию, которая не может быть отражена в онтологии посредством построения иерархии понятий и установки различных отношений между понятиями. В качестве примера аксиомы

можно привести следующее высказывание: «Если X смертен, то X когда-нибудь умрет». Аксиомы позволяют в дальнейшем осуществлять умозаключения в рамках онтологии. Они могут снабжать исследователей информацией о правилах, позволяющих автоматически добавлять информацию. Аксиомы могут также представлять собой ограничения, накладываемые на какие-либо отношения, делающие возможным выведение умозаключений. Приведем несколько примеров таких ограничений. Понятийные ограничения указывают на то, какой тип понятий может выражать данное отношение (например, свойство Цвет может выражаться только понятиями категории Цвет). Примером числовых ограничений является утверждение того, что для Человека количество биологических родителей равно 2. Количество и степень детализации аксиом обычно зависят от типа онтологии, о чем будет подробнее сказано далее.

Наряду с указанными элементами онтологии в нее также входят так называемые «экземпляры». В литературе они могут выступать также под названиями:

- конкретные экземпляры;
- инстанции;
- индивидуальные экземпляры.

Экземпляры – это отдельные представители класса сущностей или явлений, то есть конкретные элементы какой-либо категории (например, экземпляром класса Человек будет королева Виктория).

Составляющие онтологии подчиняются своеобразной иерархии. На нижнем уровне этой иерархической лестницы находятся экземпляры, конкретные индивиды, выше идут понятия, то есть категории. На уровень выше располагаются отношения между этими понятиями, а обобщающей и связующей является ступень правил или аксиом.

Как упомянуто в работах [3] и [5], «термину «онтология» удовлетворяет широкий спектр структур, представляющих знания о той или иной предметной области». Так к онтологиям можно отнести ряд структур, отличающихся разной степенью формализованности:

- глоссарий;
- простая таксономия;

- тезаурус (таксономия с терминами);
- понятийная структура с произвольным набором отношений;
- полностью аксиоматизированная теория.

Однако в этих структурах не всегда представлены все составляющие онтологии, которые описывались в данном разделе.

1.3. Классификация онтологий

Онтологии сильно различаются по ряду параметров, и исследователи выделяют различные основания для их классификации. Так Э. Хови [20] говорит, что онтологии различаются в зависимости от набора элементов, содержащихся в них, а также типов вводимых отношений. Он выделяет так называемые «терминологические онтологии» и «настоящие онтологии». Под первыми Э.Хови [20] понимает онтологии, включающие сущности, явления, свойства, связи предметной области и объединяющие их структурные отношения. «Настоящие» же онтологии включают в себя также дефиниционные отношения и отношения дополнительной информации. Наряду с этим в них входят аксиомы, определяющие взаимозависимости между отношениями и понятиями.

Э.Хови выстраивает подробную классификацию различных характеристик онтологий. Он упоминает, что основными параметрами могут быть: форма (то, как формируется онтология), содержание, а также средства использования онтологии.

Можно упомянуть о том, что существует разбиение онтологий по количеству и качеству понятий, включаемых в них. Онтологии верхней зоны обычно насчитывают примерно 100-500 концептов. В них включены наиболее абстрактные категории, обладающие свойством универсальности. Они являются базовым разбиением наблюдаемой действительности на категории. Обычно они строятся теоретиками, философами. Зачастую концепты даже не лексикализуются. Составление аксиом в данном типе онтологий с высоким уровнем обобщения достаточно сложно и требует некоторого воображения. Преимуществом таких онтологий является возможность их использования во многих областях и даже во многих языках. Для данного рода онтологий характерен ограниченный набор обобщенных отношений, которые можно отнести к базовым (таких как родовидовые отношения, отношения часть-целое и ассоциативные отношения). В таких онтологиях типичными на верхнем уровне разбиения являются такие понятия, как:

- сущность;
- явление;
- объект;

- процесс;
- роль [3].

Однако этот типичный набор может быть представлен в усеченном виде, например, в онтологии MicroKosmos, разработанной С. Ниренбургом и В. Раскиным [26] на верхнем уровне появляются лишь три категории понятий: «объект», «процесс» и «роль». При этом авторы претендуют на универсальность этого разбиения для онтологической семантики в целом и, таким образом, для всех онтологий верхнего уровня.

Другим типом являются онтологии средней зоны, здесь элементов обычно уже больше (500 – 100000 концептов). Они представляют мир в целом и в общем случае это неаксиоматизированная область. Сложность заключается в том, что для данного вида онтологий требуется выводить слишком большое количество аксиом. Обычно выходом является использование методов автоматизированного вывода аксиом из уже существующих онтологий. Построением онтологий этого уровня чаще всего занимаются когнитологи и лингвисты.

Онтологии нижней зоны или так называемые онтологии предметной области наиболее обширные, обычно они насчитывают около 200 – 2 000 концептов. Они описывают конкретные предметные области с их спецификой. При этом круг решаемых задач и вопросов, на которые онтология отвечает, ограничен выбранной областью. Для данного типа онтологий характерно наличие отношений, специфичных для конкретной области [3]. Это высокоаксиоматизированная зона, то есть для нее возможно построение большого количества аксиом и правил. В большинстве случаев этот тип онтологий строится экспертами области знания или при их содействии. В связи с большой спецификой каждой отдельной предметной онтологии ее повторное использование зачастую возможно только в рамках предметной области.

Наряду с описанным делением все онтологии могут быть разделены на глубинные и поверхностные. Поверхностные онтологии строятся на поверхностной семантике, они определяют понятия через значения слов. Однако здесь возникает проблема, какое количество смыслов выделять для каждого слова. Глубинные же онтологии используют глубинную семантику.

Б. Й. Вилинга и А. Т. Шрайбер выделяют два измерения для оценки онтологий: «объем и тип структуры концептуализации и предмет концептуализации» [34]. Однако, как указывает Н.Гуарино [18], критерий для выделения в рамках первого измерения информационных, терминологических онтологий и онтологий, моделирующих знания, не отличается четкостью. Проще разделить онтологии по степени детальности, используемой для характеристики концептуализации. Очень детализированная онтология подробно специализирует подразумеваемую концептуализацию, но платой за это оказывается более громоздкий язык, который может быть сложно применять на практике. Простая же онтология может развиваться с допущением каких-то скрытых условий, которые подразумеваются создателями, и ее могут использовать те, кто уже договорился о лежащей в основе концептуализации (и осознает эти допущения). Можно различать «отсылочные (также называемые офф-лайн) онтологии» (reference ontologies) и «осуществляемые (совместно используемые, он-лайн) онтологии» (implemented (shareable) ontologies). Несложная структура, описывающая, например, лексикон, может помещаться он-лайн, в то время как замысловатые теории, определяющие значение терминов из лексикона, могут находиться офф-лайн.

С точки зрения предмета концептуализации исследователи выделяют прикладные онтологии, онтологии области знания, общие (родовые) онтологии и репрезентационные онтологии (речь идет об онтологиях метауровня, включающих в себя репрезентационные первоэлементы) [18].

Онтологии могут быть также разделены на одноязычные и многоязычные. Уже существует ряд онтологий, ориентированных на представление знаний на нескольких языках, например, EuroWordNet, MikroKosmos и некоторые другие. Сложность создания таких онтологий обычно заключается в том, что возможно наличие различий в понятийных систем разных языков.

В рамках работы [3] также выделяется особый тип онтологий – лексические (или лингвистические). Отличительным свойством таких онтологий является «фиксация в одном ресурсе (лексикализованных) понятий (слов) вместе с их языковыми свойствами»[3]. Такие онтологии тесно взаимосвязаны с семантикой грамматических элементов (слов, именных групп и др.). Основным источником понятий в онтологиях данного типа являются значения языковых единиц. Их также отличает своеобразный

набор отношений, обычно свойственный для языковых элементов: синонимия, гипонимия, меронимия, а также ряд других. К лингвистическим онтологиям авторы [3] относят WordNet, MikroKosmos, Sensus, РуТез и другие. Круг задач, решаемых такими онтологиями, тесно взаимосвязан с обработкой естественного языка.

С.А.Коваль [8] предлагает различать безэкземплярные и экземплярные онтологии. Как понятно из названия данного типа онтологий, безэкземплярные онтологии отличаются отсутствие конкретных экземпляров. На нижних уровнях иерархии таких онтологий находятся не конкретные экземпляры, а понятия. Эта особенность онтологии накладывает некоторый отпечаток и на вводимые в данной онтологии отношения.

Таким образом, существует множество подразделений онтологий, но эти классификации не всегда бывают достаточно четкими и последовательными.

Создаваемая нами онтология относилась к онтологиям нижней зоны и описывала конкретную предметную подобласть. В связи с этим для нашей онтологии были характерны некоторые отношения, которые являются специфичными для данной области, и мала вероятность, что могут использоваться в других областях. Она также являлась безэкземлярной, что наложило свой отпечаток на установление отношений между элементами, однако об этом будет упомянуто подробнее в разделе 2.5. На данном этапе создания можно говорить об одноязычности нашей онтологии, хотя для некоторых элементов уже была сделана попытка связать русские языковые выражения с английскими.

1.4. Методы построения онтологий

1.4.1. Принятие исходных решений

Определившись с понятием «онтологии» и ее структурой, перейдем к обсуждению методов их построения. Говоря в общих чертах, для создания онтологии надо сначала перечислить категории, обозначающие сущности или явления в моделируемой области. Затем следует связать эти категории определенными отношениями. И на последнем шаге надо соотнести категориям набор конкретных экземпляров. Но это лишь общий, упрощенный алгоритм, а в реальности этот процесс противоречив и рождает много дискуссий.

Так, ряд решений должен быть принят уже на начальных этапах создания онтологий. Надо определить, создавать новый элемент и должен ли этот элемент быть включен в структуру. Следует понять, где по отношению к другим объектам должен располагаться вновь создаваемый, должен ли он быть видом какого-либо класса или же сам представляет собой родовой термин. Помощь в этом может оказать формулировка особых, уникальных свойств термина, то есть его отличительных характеристик. При этом не следует смешивать свойства понятия и его отличительные признаки.

При формировании онтологий могут привлекаться специалисты различных областей, и для каждой области есть свои базовые методы работы. Так, философы используют абстракцию и комбинирование свойств, например, по Аристотелю следует выделять понятия как атомарные понятия, строящиеся из набора дифференциальных признаков. Когнитивисты склонны полагаться на интуитивные отличия. Например, Э.Рош [29] полагает, что функциональным предназначением классов является предоставление максимума информации при минимуме когнитивных усилий. Считая, что люди склонны формировать классы на основе прототипов, она вводить такие классы предлагает в дополнение к родовидовым иерархическим системам или вместо них. Специалисты в компьютерной области оперируют логическими теориями и используют структуры, построенные на умозаключениях. Лингвисты уделяют большое внимание описаниям межъязыковых соответствий. Общим же методом для всех является использование классификаций.

Однако при реализации упомянутых ранее в общих чертах методах построения онтологии может возникнуть ряд проблем. Так при составлении классификаций существует возможность появления неоднозначностей в зависимости от того, какому дифференциальному признаку отдавать предпочтение. Можно привести пример разбиения понятия «человек», предлагаемый Э.Хови [20]. С одной стороны, можно сначала использовать отличительный признак «пол», выделив понятия «лица мужского пола» и «лица женского пола», а затем «возраст», выделив «мужчина» и «мальчик», а также «женщина» и «девочка». Но, с другой стороны, в равной степени правомерным будет разделить понятие «человек» на «взрослые» и «дети», а потом уже выделить пару «мужчина» и «женщина» и пару «мальчик» и «девочка». В данном случае проблема заключается в неопределенности иерархии отличительных признаков.

Выбор того или иного решения упомянутых ранее проблем порой достаточно сложно обосновать. В данном случае нет авторитетов, это может быть данью традиции, общественным соглашением или просто подстройка под определенные задачи. Самым важным здесь является выбрать какой-либо подход и придерживаться его на протяжении всей работы, поэтому стоит задуматься над рядом проблем и возможными универсальными методами их устранения еще до начала принятия решений.

При создании концептов исследователи сталкиваются еще с одной принципиальной проблемой, так называемой «проблемой примитивности». Э.Хови [20] указывает на то, что почти вся семантика и представление знаний основываются на композиционной гипотезе: можно определить ограниченный набор единичных сущностей («атомов»), а все остальные («молекулы») представлять как комбинацию (композицию) атомов. При этом возникает вопрос: как много таких «атомов» необходимо? В современных исследованиях можно найти два различных подхода: экономный и неэкономный.

Экономный подход призывает создавать малое количество элементарных концептов, семантически простых, с помощью которых можно объяснить значение более сложных понятий. При таком положении вещей легко обнаружить связанность понятий, просто определять и осуществлять умозаключения, однако достаточно сложно составлять сложные значения. Примером такой точки зрения является исследование А.Вежбицкой [36], автора теории семантических примитивов и основанного на ней

естественного метаязыка описания семантики. Семантические примитивы – это лексические единицы, выражающие элементарные, базовые значения. Количество таких единиц не превышает 100, и их значения универсальны для всех языков.

Неэкономный подход позволяет создавать любое количество индивидуальных сущностей – столько, сколько захочется создателю онтологии. Это количество может варьироваться от 10 до 100 000 и более. Такая точка зрения свойственна, например, создателям WordNet [25]. Здесь затруднительно определять связанность понятий, сложно работать с умозаключениями. Данный подход сопровождается, по существу, отказом от композиционной гипотезы, но это влечет за собой и преимущество: отсутствие необходимости составлять сложные значения.

В данной проблеме нет единственно верного решения: все зависит от того, как много высказываний необходимо или как сложна исследуемая предметная область. Подытоживая, следует сказать, что современная практика показывает, что экономного полхода придерживаются в основном формалисты, а пользователи склоняются к неэкономному.

Стоит упомянуть, что не всегда построение онтологии проходит очевидным образом, не всегда легко собрать понятия, выделить дифференциальные признаки. Здесь также существует несколько вариантов действий, зависящих от конкретных задач и исходного материала. Существует возможность сбора элементов для онтологии напрямую. При таком подходе сначала собираются и классифицируются понятия, подбираются слова, затем проводится соответствие между понятиями и лексиконом. Проблемой является близкое, но не идентичное пересечение значений в словах разных языков. Вторым вариантом является использование микротеорий. При таком подходе сначала надо понять явление, затем формируются примитивы теории, элементы лексикона определяются с точки зрения примитивов, после этого лексикон усложняется. Проблема данного метода заключается в выборе микротеории, а также в необходимости отдельной теории для каждого комплекса значений. Противопоставление этих двух точек зрения можно проиллюстрировать на предложенном Э.Хови [20] примере, где рассматривается несколько подходов к трактовке понятия «цвет» и различных обозначений цветов. Можно напрямую собрать слова, обозначающие цвет и попробовать их как-нибудь классифицировать. А можно

применить микротеоретизирование, понять, что представляет собой явление «цвет». С физической точки зрения цвета задаются длиной волны и интенсивностью, теперь с помощью этих параметров мы можем определить все конкретные значения цвета. Может быть применен и нейрофизиологический поход, представляющий все цвета с точки зрения восприятия их с помощью трех рецепторов.

Достаточно ясно, что в основании онтологии должны лежать понятия, но возникает вопрос: как при построении определить понятия, основываясь на словах? Обычно связующим звеном становится категория значения (лексико-семантического варианта). Слова существуют в рамках одного языка, значения же независимы от конкретного языка. Э.Хови [20] предлагает формальную процедуру выявления понятий на базе совокупности слов. Он предлагает алгоритм перехода от слов к значениям, а затем к понятиям:

1. **Инициализация**: Для данного слова соберите несколько десятков предложений, содержащих его. Подберите определения из различных словарей.
2. Расположите значения слова в предварительные, грубо схожие группы
3. **Процесс дифференциации:** Начните строить дерево, расположив все группы в корне
4. Рассматривая все группы, определите группу, наиболее отличную от других:
    1. Если вы можете найти одну ясно выделяющуюся группу, выпишите ее наиболее яркое отличие в явной форме – оно послужит отличительным признаком и будет формализовано в виде аксиомы.
    2. Если отличия, по которым можно далее подразделить группу, не обнаруживаются, остановите работу с этой ветвью и перейдите на другую ветвь
    3. Если обнаруживается несколько отличий, позволяющих подразделить группу несколькими равнозначными способами, также прекратите работу с этой ветвью и перейдите на другую ветвь.
5. Создайте в древесной структуре две новые ветви, расположите новую группу под одной ветвью, а остальные под другой.

6. Повторите действия с шага 4, исследуя по отдельности группу/группы под каждой ветвью

7. **Формирование понятий**: Когда ветвление прекращается, конечным результатом является дерево все более дробных отличительных признаков, которые в явном виде перечислены на каждом уровне дерева. Каждый лист становится отдельным понятием, далее не делимым в настоящей задаче (приложении, предметной области). Каждое отличие должно быть формализовано в виде аксиомы, которая срабатывает для ветви, с которой ассоциируется.

8. **Добавление понятия в онтологию**: Начиная с вершины, пройдите каждый узел с ветвлением. Имеют ли уже созданная и добавляемая ветвь примерно одно и то же значение?

1. Если так, объедините их в онтологии в подходящем узле и остановите прохождение этой ветви

2. Если нет, разделите дерево и повторите шаг 8 для каждой ветви. Повторяйте, пока не дойдете до конца.

Исследования показывают, что обычно понятий оказывается меньше, чем значений-смыслов.

Однако существуют и другие источники нахождения новых понятий. Для этих целей можно использовать уже существующие онтологии и различные списки, словари и тезаурусы. На данном этапе также может помочь автоматическое выявление понятий путем кластеризации слов (о ней подробнее в раздел 1.4.2).

Достаточно много внимания методике создания онтологий уделяется в статье Н. Ной и Д. МакГиннес [27]. Основополагающие правила разработки онтологии авторы формулируют следующим образом:

1) Не существует единственного правильного способа моделирования предметной области – всегда существуют жизнеспособные альтернативы. Лучшее решение почти всегда зависит от предполагаемого приложения и ожидаемых расширений.

2) Разработка онтологии – это обязательно итеративный процесс.

Под итеративным процессом понимается неоднократный проход по онтологии с целью ее уточнения, то есть на начальном этапе строится черновой вариант. Затем мы проверяем и уточняем составленную онтологию, добавляя детали, возможно, частично или даже полностью пересматривая начальную онтологию.

3) Элементы онтологии должны быть близки к объектам (физическим или логическим) и отношениям в интересующей вас предметной области. Скорее всего, они соответствуют существительным (объекты) или глаголам (отношения) в предложениях, которые описывают вашу предметную область.

Все же следует упомянуть, что каким бы образом не решались ранее упомянутые проблемы, эти решения должны быть последовательными. Таким образом, при построении онтологии нам нужны определенные ориентиры для принятия решений. Целый ряд решений может приниматься на основе практической цели построения онтологии. Однако можно сформулировать и универсальные требования к онтологиям, не зависящие от конкретной задачи. Так, общая структура онтологии должна быть понятной и должна существовать возможность ее многократного использования. Как подчеркивает Н.Гуарино [18], онтология должна быть когнитивно прозрачной. Ряд требований к онтологии можно найти в работе С. Ниренбурга и В. Раскина [26]:

- Ясность: онтология должна быть ясной и легко передавать подразумеваемый смысл. Она должна быть объективной;
- Последовательность: в ней должны содержаться утверждения, которые не противоречат друг другу, иерархии понятий, связывающим их отношениям, экземплярам.
- Возможность расширения: наличие возможности введения новых элементов без пересмотра остальных элементов;
- Минимальная степень специализации онтологии: нежелательность полного подчинения онтологии конкретной задаче, что может осложнить ее последующее использование в других задачах.

Нельзя утверждать, что этот список требований к онтологиям является исчерпывающим, но он может помочь при принятии тех или иных решений, касающихся структуры онтологии.

Существуют и более формализованные и подробные описания стандартов для онтологий. Так, можно привести пример инициативы EAGLES - Expert Advisory Group on Language Engineering Standards (http://www.ilc.cnr.it/EAGLES96/home.html), действовавшую с 1993 года по 1996, и далее сменившую ее ISLE - International Standard for Language Engineering (http://www.mpi.nl/ISLE/). Целью данных проектов является разработка универсальных стандартов и рекомендаций для создания языковых ресурсов и программ, обрабатывающих естественный язык. Созданные в рамках EAGLES стандарты уже давно стали общепринятыми и широко распространенными.

Существует множество стандартов по языкам представления онтологий, а также по общим правилам создания лингвистических ресурсов для последующей компьютерной обработки. Список таких стандартов (в том числе и для онтологий) можно, например, найти на сайте The Language Technology Resource Center (http://flrc.mitre.org/References/Standards/). Использование стандартов должно стать залогом того, что созданный ресурс будет легко внедряться в уже существующие и учитывать все особенности технологии.

1.4.2. Автоматические методы построения онтологий

Как уже было упомянуто ранее, сейчас развитие онтологий начинает приобретать более массовый характер, и в настоящее время в этой области есть ряд масштабных разработок. Существует большое количество различных списков и баз данных, но возникает вопрос, как гарантировать их соответствие текущему положению вещей, как быть уверенным, что они точны и полны, а также как обеспечить достаточную детальность представляемых данных. В связи с тем, что мир очень быстро изменяется, идет развитие новых отраслей, существующие онтологии требуют постоянного пополнения и усовершенствования. На данном этапе появляются идеи использования автоматических и полуавтоматических методов для не только обновления онтологий, но даже для их создания.

Существует ряд методов расширения онтологий, которые специфичны для онтологий разных зон. Для расширения верхней, наиболее общей зоны необходимо подробное теоретизирование, после чего можно приступать к построению понятий и

аксиом. Для онтологий средней зоны, которые отличаются большим количеством понятий, сбор понятий может выполняться автоматически с помощью кластеризации. В процессе обработки большого количества информации происходит сбор понятий и разбиение их по классам на основании каких-то общих характеристик. Существует целый ряд методов по увеличению точности извлечения семантически связанных семей понятий. При таком анализе в дальнейшем возможно также устанавливать перекрестные ссылки внутри онтологии. Однако для онтологии важно знать не только то, что понятия взаимосвязаны, но, и то, как именно они взаимосвязаны. Для выявления таких отношений между понятиями также могут быть использованы автоматические методы просмотра и анализа различных текстов, например, как предлагают авторы работы [21], можно извлекать данную информацию из словарных определений. Это обусловлено тем, что существует ограниченный набор фразовых моделей, вводящих определяющее, по характеру которых можно сформулировать тип связей между понятиями и ввести эти данные в онтологию. Таким образом, можно выбрать словарные статьи, провести их разбор, выявить семантические толкования для слов и начальных конструкций определений. Данная идея пополнения онтологии встречается и в работе [15], где такой подход позволяет выявить отношения между элементами онтологии с помощью анализа корпуса и поиска моделей, соответствующих какому-либо роду отношений.

Выявление же аксиоматических знаний, правил может быть произведено на основании Интернета. Э.Хови указывает на то, что сбор отдельных примеров происходит за счет изучения большой базы ресурсов, однако вначале формулируется ряд параметров, указывающих, какие именно экземпляры нам нужны [20]. Это можно пояснить на конкретном примере: если нам нужно собрать названия столиц, то можно, например, задать условия поиска с моделью "X – столица ...". При этом можно использовать не одну такую модель, а несколько. Таким образом, нахождение отдельных экземпляров будет сводиться к анализу текстов для выявления примеров с данными структурами.

Такие шаблоны могут формироваться как вручную, так и автоматически с помощью самообучающихся программ.

Наряду с поиском отдельных экземпляров важным является установление различных связей между элементами онтологии, классами. В целом авторы работы [28] предлагают разделить все методы извлечения отношений между элементами онтологии на два класса: подходы, основанные на использовании шаблонов, и методы, использующие кластеризацию. При использовании методов на основе шаблонов исследователи ищут языковые модели, которые указывают на какой-либо тип отношений между классами. В большинстве случаев осуществляется поиск родовидовых отношений и отношений «часть-целое».

При наличии базового перечня категорий можно наращивать их число, обращаясь к корпусу текстов. В корпусе выделяются кластеры близких по значению элементов, затем оценивается теснота связи между элементами, далее каждому кластеру приписывается имя, которое ассоциируется с категорией, включаемой в онтологию (подробнее см., например, [23], [30]). Таким образом, пополнение онтологии новыми элементами происходит автоматически и также автоматически определяется место нового элемента в иерархии категорий.

Как упоминалось ранее, в онтологии присутствуют не только классы и отдельные экземпляры, но и аксиомы. Они вносят большой вклад в усовершенствование анализа информации, дают возможность компьютеру дополнить текст дополнительными знаниями, которые для нас кажутся тривиальными. Аксиомы сообщают компьютеру, например, о том, что два высказывания имеют один и тот же смысл или же один факт влечет за собой наличие другого. В работе [12] рассматривается метод автоматического извлечения аксиом из текстов и их дальнейшая классификация и проверка на релевантность. В качестве основной проблемы автоматического выявления аксиом авторы упоминают построение ошибочных предположений или слишком общих, а также проблему определения симметричности аксиом. Они берут за основу гипотезу о схожести значений при схожести контекстов встречаемости, дополняя их своими ограничениями. Здесь ключевым становится определение методов вычисления близости контекстов и близости понятий. По утверждению авторов их методика выявления аксиом позволяет извлекать релевантные аксиомы с маленьким процентом ошибок.

В большинстве случаев проблемой автоматического извлечения становится большое количество «шума», который надо эффективно отсеивать. В связи с этим иногда наряду с автоматическими методами используют последующую ручную обработку полученного материала для получения данных большей точности.

В работе [28] упоминаются общие требования, предъявляемые к системам автоматического извлечения данных для онтологий:

- Минимальный контроль – сведение к минимуму или исключение вообще участие человека.
- Универсальность – применимость к различным источникам, все зависимости от их размера, области знания и т.д.
- Точность – извлеченная информация должна содержать как можно меньше ошибок.
- др.

Выполнение данных требований, возможно, позволит построить эффективную систему автоматического построения онтологий, пока же все существующие системы нуждаются в доработках и улучшениях или же успешно работают лишь применительно к замкнутым областям знания.

Однако никакая онтология не полезна в изоляции, ее преимущество как раз заключается в возможности использоваться вновь в новом окружении. Сейчас существует много различных алгоритмов состыковки нескольких уже созданных онтологий (их общее описание приводится в [18], [20]). Эти методы используются для слияния онтологий и для нахождения соответствий для слов в других языках (в данном случае онтология служит межъязыковой точкой отсчета).

Соединение двух онтологий осуществляется в несколько этапов. На начальной стадии происходит нахождение связующего звена, на основании которого можно произвести слияние. Затем срабатывает выравнивающий алгоритм, который по описанию элементов находит их место в новой структуре. Потом уже идет выверка результатов состыковки в контексте.

Как упоминалось ранее, существует целый набор алгоритмов по нахождению связующего звена для двух онтологий, стоит рассказать о них немного подробнее. Выделяют несколько различных методов обнаружения связующих звеньев ([20]):

- текстовые совпадения;
- совпадения иерархических отношений;
- совпадение форматов и данных.

Э.Хови [20] поясняет, что под текстовыми совпадениями подразумевается идентичность имен понятий (здесь также учитываются родственные слова), текстовых определений (сравнение строк, трансформация, исключение стоп-слов и др.). Иерархическое совпадение предусматривает поиск общих вышестоящих понятий, фильтрацию неоднозначностей, нахождение семантического расстояния, рассеивание семантических групп (semantic group dispersal). Совпадение форматов и данных опирается на внутрипонятийные отношения и ограничения на заполнение слотов. После отработки алгоритмов используется функция, которая учитывает результаты всех процедур и выдает общий коэффициент совпадения. Нахождение связующего звена включает в себя также процедуры валидации, в ходе которых происходит проверка с учетом иерархических связей соотносимых понятий. Эта процедура пытается найти несоответствия, порочные круги, проверить наследование свойств.

Для выявления идентичности понятий используются специально созданные критерии. Так, комплексный критерий, предложенный Н. Гуарино, проверяет сходство по нескольким параметрам:

- материал: идентичность материала, из которого сделаны экземпляры сравниваемых понятий;
- топологический: идентичность формы экземпляров сравниваемых понятий;
- морфологический: те части, из которых состоят экземпляры сравниваемых понятий;
- функциональный: использование;
- меронимический: экземпляры понятий;
- социальный: социальная роль [20].

Учитываются также возможные стандартные метонимические переносы, которые делают онтологию более гибкой и расширяют возможность нахождения близких по содержанию понятий.

Упомянутые ранее методы являются полуавтоматическими, то есть сначала автоматически генерируются варианты соответствий, а потом вручную в несколько этапов происходит соединение онтологий. Как упоминает Э.Хови [20], статистика показала, что эти процедуры обладают достаточно высокой степенью точности и дают хорошие результаты. Так, использование подобных автоматических алгоритмов выравнивания при построении онтологии SENSUS дало более 90% точности.

Побочным положительным результатом использования процедур слияния является возможность выявить ошибки и опущения в онтологиях. Слияние онтологий может использоваться также для устранения проблемы дублирования. На данном этапе развития происходит так, что соревнующиеся и взаимосвязанные коллективы формируют схожие модели областей знания, где наблюдается частичное пересечение. При такого рода дублировании неизбежно страдает и логичность. Следует признать, что некоторые пересечения конечно необходимы.

Однако Э.Хови [19] отмечает, что при слиянии онтологий может возникнуть ряд проблем, которые может быть достаточно сложно решать автоматическими методами. Так эксперты в разных областях могут отсылать к одному и тому же понятию и понимать его различным образом. Проблемы возникают и тогда, когда одно и то же слово используется для обозначения различных понятий в каждом отдельном поле. Решением такой проблемы может быть более тесная коммуникация составителей онтологий, а также использование более широких онтологий, применимых к различным областям знания.

1.5. Применение онтологий

В предыдущих разделах рассматривались определения и структура онтологий, но остался нераскрытым вопрос, зачем строят онтологии и где они применяются. Н. Ной [27] упоминает ряд способов использования онтологий:

- для совместного использования людьми или программными агентами общего понимания структуры информации;
- для возможности повторного использования знаний в предметной области;
- для того чтобы сделать допущения в предметной области явными;
- для отделения знаний в предметной области от оперативных знаний;
- для анализа знаний в предметной области.

Построение онтологии часто не является само по себе конечной целью, обычно онтологии далее используются другими программами для решения практических целей. На данном этапе развития науки существует ряд задач, где применение онтологий может дать хорошие результаты. Однако сейчас лишь малое количество приложений на естественном языке включают в себя онтологические базы, откуда черпаются знания об окружающей действительности. С. Ниренбург и В. Раскин [26] говорят о возможности использования онтологий в:

- машинном переводе;
- вопросно-ответных системах;
- информационном поиске;
- системах извлечения знаний;
- общих системах ведения диалога между компьютером и человеком;
- системах понимания языка (автоматическое реферирование текста, рубрикация и др.)

Можно упомянуть также системы расширенного консультирования, которые включают в себя несколько уровней работы с информацией и строятся на базе другие приложений.

В искусственном интеллекте онтологии используются для формальной спецификации понятий и отношений, которые характеризуют определенную область знаний. Поскольку компьютер не может понимать, как человек, положение вещей в мире, ему необходимо представление всей информации в формальном виде. Таким образом, онтологии служат своеобразной моделью окружающего мира, а их структура такова, что легко поддаются машинной обработке и анализу. Онтологии снабжают систему сведениями о хорошо описанной семантике заданных слов и указывают иерархическое строение области, взаимосвязь элементов. Все это позволяет компьютерным программам при помощи онтологий делать умозаключения из представленной информации и манипулировать ими.

Онтологии используются также при построении корпуса определений, служащего справочным материалом. В дальнейшем результаты этой работы могут использоваться для сложных процедур обработки естественного языка, например, в снятии омонимии на основе контекста. Онтологии могут использоваться для вывода умозаключений, необходимых для понимания текстов на глубинно-семантическом уровне, что требуется для высококачественного машинного перевода и может служить базой для расширения и уточнения информационного поиска. Глубокий анализ текста необходим и для систем автоматического реферирования. Стоит упомянуть, что также онтологии могут способствовать систематизации понятий. На базе онтологий может осуществляться автоматическое аннотирование и разбор текстов, которое в дальнейшем может использоваться в первую очередь в информационном поиске, а также при различных видах анализа информации.

Приведем некоторые примеры существующих систем, содержащих онтологические приложения. В сфере информационного поиска заслуживает упоминания европейский исследовательский проект под названием CROSSMARC (описание данной системы приводится в работе [22]). Участники этого проекта делают упор на необходимости широкого использования онтологий для разделения отраслевых и общепонятийных знаний, считая, что это облегчит извлечение информации из

различных источников, сузит поисковые запросы и улучшит качество выдаваемых результатов. Эта задача оказывается смежной с задачей автоматической рубрикации текстов, в ходе которой производится распределение текстов по рубрикам на основе автоматических методов и использования онтологий.

В области машинного перевода известна система OntoLearn (описание данной системы приводится в работе [22]), используемая при переводе многословных терминов с английского языка на итальянский. Система автоматически вычленяет и строит предметные онтологии. Промежуточные онтологические построения используются для прямого машинного перевода. Можно привести также пример системы машинного перевода, разработанной в Университете Sains в Малайзии [24]. Она осуществляет снятие неоднозначности со слов, используя тексты определений и структурную информацию из онтологий.

IAMTC (Interlingual Annotation of Multilingual Text Corpora) можно отнести к системам понимания языка. Этот многосторонний проект занимается аннотацией шести больших параллельных корпусов с целью извлечения межъязыковых соответствий. Система использует 110 000 записей онтологии OMEGA для частеречной разметки и дальнейшего анализа естественного языка. Для понимания естественного языка могут использоваться также аксиомы и умозаключения, содержащиеся в онтологиях, помогает и большой набор отдельных примеров, экземпляров.

Как уже упоминалось ранее, онтологии могут также лежать в основе различных вопросно-ответных систем и способствовать улучшению анализа запросов и точности ответов. Можно привести пример демонстрационной вопросно-ответной системы YAWA [7]. YAWA по запросу выдает информацию о главах государств и правительств стран мира. Она обладает сведениями о том, кто занимает указанную должность в данной стране. Кроме этих связей, в нее заложены знания о соотношении названия и основной функции (глава государства и/или правительства) высших государственных должностей в отдельно взятой стране в зависимости от типа системы государственного управления. Таким образом, при выдаче ответов по запросу система учитывает заложенные в нее сведения об окружающей действительности: набор понятий, отношений между ними, ограничений на отношения и список конкретных экземпляров.

В работе [14] рассматривается целое направление работ в сфере электронной коммерции, где онтологии предоставляют классификацию товаров и услуг и обеспечивают наличие стандартизованного представления информации. Таким образом, происходит систематизация понятий области бизнеса, упорядочение их описаний. Онтологии предоставляют эффективный доступ к информации, дают возможность лучше понять данную информацию и, следовательно, произвести её более широкий и сложный анализ. Это использование онтологий является ярким примером использования онтологий «для совместного использования общего понимания структуры информации людьми или программными агентами», упоминаемого Н. Ной [27].

Еще один пример использования онтологий представляет собой система понимания языка, разработанная в НПЦ «Интелтек Плюс» [1]. Сейчас уже «создана и внедряется в Совете Федерации Федерального Собрания Российской Федерации первая очередь информационной системы "Семантический контроль текстов редактируемых документов"... Она используется специалистами Управления информационного и документационного обеспечения аппарата Совета Федерации для проверки правильности расшифровки стенограмм и проверки редакций различных типов документов на предмет их соответствия эталонным словарям и базам данных»[1]. Эта система занимается поиском несоответствий в текстах редактируемых документов на основе эталонного описания предметной области, содержащегося в онтологии. К таким несоответствиям авторы относят «ошибочные должности сотрудников организаций, ссылки на устаревшие структурные подразделения организаций, неправильные телефонные номера должностных лиц» [1]. Таким образом, они стремятся выявить «неэквивалентность факта, выявленного при анализе текста, имеющимся в базе знаний фактам». В данном случае под базой знаний понимается онтология, снабженная конкретными экземплярами. После извлечения знаний из текстов, модуль логического вывода сверяет их с данными в онтологии, проверяя наличие связей между элементами, отслеживая правильность этих связей, и таким образом подтверждает или опровергает достоверность фактов и отмечает нарушение семантических связей.

Онтологии занимают ключевую позицию во многих лингвистических комплексах, так, например, InfoMap (http://infomap.stanford.edu) широко использует

иерархические структуры в работе с текстами на естественном языке. Целью данного проекта являлось извлечение значения слов на базе их употребления в тексте. Этот материал в дальнейшем может использоваться для понимания языка и машинного перевода, а также для интеллектуального поиска, принимающего во внимание значения слов, и индексирования ([9]). В рамках данного проекта при группировке понятий и значений используется верхняя зона WordNet, а также анализ корпусов текстов путем извлечения кластеров слов автоматически. Затем эти данные используются для подбора вероятных родовых терминов в группах слов, то есть для формирования таксономий. Также возможно применение результатов данных процедур для сравнения структуры групп в параллельных текстах на разных языках, что позволяет улучшать качество машинного перевода и уточнять переводы слов в рамках многоязычной лексикографии.

Существует большое количество проектов в области медицины, использующих онтологии в своих приложениях. Так, можно привести пример проекта MuchMore (http://muchmore.dfki.de), являющего частью InfoMap, описанного ранее, занимающего разработкой методов организации информации на различных языках и в частности медицинской области знания. Их исследование основывается на использовании иерархии понятий для предметных областей, и следовательно технологиях извлечения многоязычных терминов и отношений. Их продукт помогает осуществлять поиск документов на различных языках по медицинской области знания. Медицинская область знания очень перспективна в этой сфере, так как для нее уже создано большое количество онтологий и структурированных источников знания, а также присутствует множество текстов на данной области, требующих обработки. Это тексты, описывающие карты больных, случаи заболеваний, общие описания разных болезней и многие другие. Проект MuchMore помогает выстроить взаимосвязи между всеми типами текстов в данной области. В задачи этого исследования входит:

- Сокращение «пропасти» между медицинской документацией и многоязычными данными путем автоматического извлечения дескрипторов и составления метаописаний истории болезней для последующего использования в других источниках. Организация информации в онтологии помогает в дальнейшем быстро строить экспертные системы и приложения для работы с данными.

- Устранение языкового барьера при поиске информации. Использование онтологий позволяет эффективно искать информацию на нескольких языках, тем самым значительно облегчая работу специалистов. Также это дает возможность сравнивать описания аналогичных случаев на разных языках и проводить более содержательные исследования.

В рамках работы [11] описывается еще одно перспективное применение онтологий, их использование при семантической разметке текста. Учет семантических категорий, описанных в онтологии, позволяет сделать разметку корпусов более точной, уменьшить неоднозначность, так как в шаблоны, по которым производится разметка, связаны в категориями в онтологии. Такая семантическая разметка в дальнейшем позволяет проводить семантический анализ текста, различные статистические исследования, извлекать межъязыковые соответствия.

Еще одним современным проектом, широко использующим онтологии является компания Онтос (http://www.ontos.com/ru/company/about_us.php), разрабатывающая различные семантические технологии. При помощи их систем, основанных на обработке текстов на естественном языке (NLP), пользователь может генерировать и хранить релевантные знания, необходимые для различных задач. Данные системы ориентированы на пользователя, которому надо обрабатывать большие массивы информации, извлекать структурированную информацию. Для решения данных задач возможно использование продуктов Онтоса, обеспечивающих автоматическую обработку необходимых неструктурированных данных и получения прямого доступа к аналитическим (обработанным) данным. Как упоминается на сайте компании, их системы успешно решают следующие задачи:

- Поддержка принятия решений при проведении исследований
- Визуализация информации с помощью семантических сетей
- Автоматическая генерация семантических аннотаций из неструктурированного текста
- Дайджестирование больших документов на базе их семантического содержания
- Резюмирование больших объемов аннотированного текста
- Поддержка мета-данных в соответствии со стандартами RDF/OWL

- Семантический поиск с применением технологии триплетов (Объект — Отношение — Объект)

Так, один из продуктов компании, OntosMiner (http://www.ontos.com/ru/products/ontos_miner.php), анализирует текст на естественном языке, используя онтологии м семантические правила. Результатом работы данной системы становится распознавание объектов и связей между ними и добавление их как аннотации к соответствующим фрагментам текста.

Еще одним аналогичным проектом является RCO (Russian Context Optimizer) (http://www.rco.ru/). С помощью современных технологий исследователи строят онтологии, семантические представления. Продукты и технологии RCO позволяют решать такие прикладные задачи как составление содержательного портрета текста, извлечение именованных объектов, связей и фактов из массивов неструктурированных данных, анализ тональности текста, выявление заимствований, обнаружение дубликатов. Использование онтологий помогает при поиске заранее неизвестной информации, относящейся к некоторой теме, позволяя выдать пользователю возможные “подсказки” для уточнения запроса. Также онтологии служат основой для решения различных аналитических задач, позволяя исследовать окружение выбранного объекта, находить цепочки и группы связности во множестве объектов.

Также в этой связи стоит упомянуть проект «Галактика ZOOM» (http://www.galaktika-zoom.ru/). Эта система предоставляет различные возможности для специалистов разных уровней: руководителей, аналитиков, маркетологов, специалистов по PR, сотрудников служб безопасности. Как упоминается на сайте, их разработки могут использоваться для поиска информации, выявления сути текста, сравнения документов и исследования документов с учетом динамики во времени.

Стоит упомянуть, что онтологии получили широкое распространения и для моделирования организационной структуры предприятий. Как упоминается в обзоре [2], онтологическое представление знаний о субъектах экономической деятельности, которые входят в состав какой-либо системы, можно использовать для объединения их информационных ресурсов в единое информационное пространство. Онтология предприятия включает в себя организационную онтологию, описывающую организационно-функциональную структуру предприятия: состав штатного расписания

(работники, администрация, обслуживающий персонал), партнеры, ресурсы и т. п. и отношения между ними, а также онтологию по технологиям, описывающую терминологию технологий. Разработанные онтологии позволяют сотрудникам одной отрасли или корпорации использовать общую терминологию и избежать взаимных недоразумений, которые могут усложнить сотрудничество и привести к серьезным убыткам.

В качестве примера практического использования онтологических моделей технологий можно привести систему ONTOLOGIC (http://www.intertech.ru/Production/sol+tech.asp), предназначенную для создания и поддержки распределенных систем нормативно-справочной информации (НСИ), ведения словарей, справочников и классификаторов и поддержки системы кодирования объектов учета. Онтология обеспечивает непротиворечивое накопление любого количества информации в стандартной структуре классификации. Такой подход гарантирует однозначную идентификацию ресурсов независимо от различных трактовок их наименований разными производителями. При использовании данной системы осуществляется эффективный контроль и верификация данных, проверки корректности, полноты и непротиворечивости данных как на этапе анализа и нормализации существующих данных, так и при занесении новых элементов данных.

1.6. Языки представления (Semantic web)

Как упоминалось ранее, на данном этапе развития технологий появилась идея расширения существующей ныне World Wide за счет добавления к уже существующим документам семантической информации. Это расширение Интернет ресурсов получило название Semantic Web. Создатель WWW Тим Бернерс-Ли определяет Semantic Web как «расширение существующей WWW, в котором информации придается четко определенное значение, позволяющее людям и компьютерам сотрудничать» (цит. по [3]). Майкл Ушолд [3] поясняет, что «точного определения SW нет, но ее основная черта – наличие семантики, доступной для обработки машинами (программными агентами Сети) при взаимодействии друг с другом». Semantic Web дает возможность компьютерам обрабатывать гипертекстовые страницы, предоставляющие информацию людям, и автоматически получать данные.

Можно привести несложный пример такой обработки, упомянутый А.Сварцем [32]: система оценки рейтинга книги. Можно добавлять баллы к конкретной книге X, если в тексте отзывов встречается фраза "Мне понравилась X", но есть и другие способы выражения того, что это хорошая книга и читатель оценил ее по достоинству. Однако компьютеру не понять, что выражает одобрительную оценку книги, для этого ему нужны данные, которые бы явно указали на это.

На решение этих и многих других задач и направлено создание такой системы, как Semantic Web. Использование стандартизованных языков позволяет точно описывать содержание, лежащее в основе HTML страниц, а также устанавливать скрытую информацию. Так, описанию могут подвергаться изображения и видеоматериалы, содержащиеся на сайтах или же какие-либо недоступные при простом чтении страницы данные.

Любой элемент в Semantic Web может получать идентификатор, называемый "Uniform Resource Identifier" или URI. Используется единообразная система идентификаторов, и каждый идентифицированный элемент рассматривается как ресурс, получивший определенное имя. Если что-то имеет URI, можно сказать, что оно находится в сети. Именно URI является основой и связующим звеном Semantic Web, и

они не подлежат перемещению. URI децентрализованы, они напрямую не контролируются никакой организацией.

В курсе лекций В. Д. Соловьева и соавторов [3] приводится структура (рис. 1), отражающая строение Semantic Web, которая у Т. Бернерса-Ли получила название «слоеный пирог». Данную структуру отличает большое разнообразие элементов:

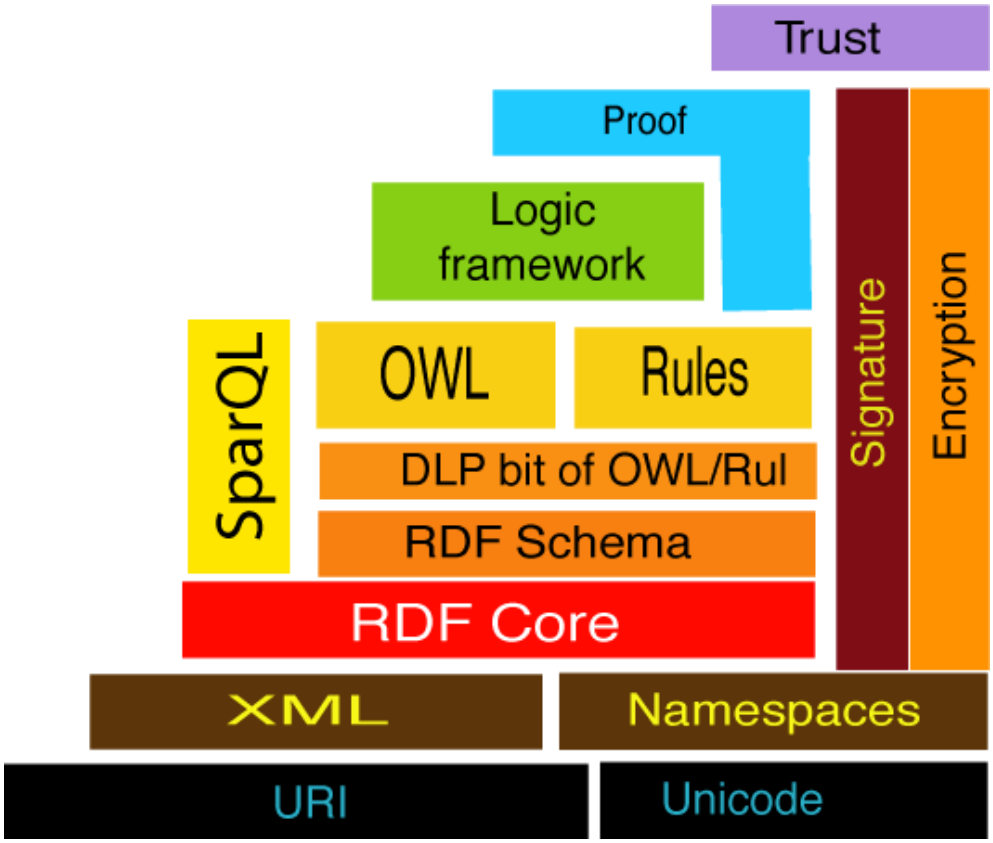

Рис. 1

Простой идентификации ресурса недостаточно, чтобы понять, как он может использоваться. Таким образом, URI бесполезны, если мы не опишем, что они значат. Для этого и нужны онтологии и так называемые схемы, они являются неотъемлемой частью Semantic Web и становятся основой для семантической разметки. В рамках Semantic Web онтологии занимают центральное положение, что можно увидеть на рис. 2, они задают отношения между понятиями и определяют логические правила для

рассуждений о них. Таким образом, компьютер может понимать смысл данных, обращаясь к онтологиям за требующейся информацией.

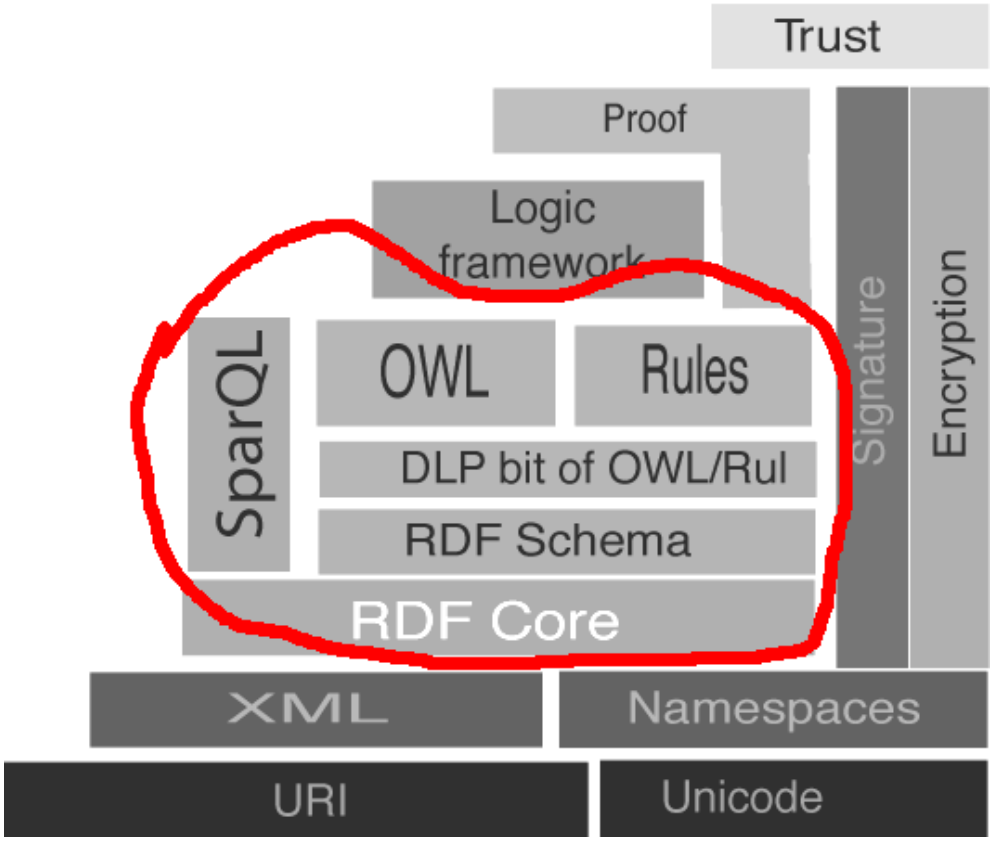

Рис. 2

Основные этапы развития языков представления онтологий в Semantic Web освещены в цикле лекций Б.В.Доброва и соавторов [3]. Изначально для описания Semantic Web использовался Extensible Markup Language (XML). Этот язык является простым инструментом создания документов с универсальными средствами описания. Он позволяет каждому описывать свою структуру, формат документа, в том числе различные виды разметки. Введенная с помощью XML «интеллектуальная» разметка поддается компьютерной обработке и может интерпретироваться особым образом.

Однако недостаточно просто описать элементы Semantic Web, надо, чтобы компьютер мог понимать и анализировать составленные нами высказывания. Больше возможностей для обработки высказывания предоставляет формализм Resource

Description Framework (RDF). Выражение RDF имеет структуру простого предложения, с единственным отличием — все слова являются URI. Каждое высказывание RDF представляет собой тройку (S,P,O), где S – субъект, P – предикат и O – объект (рис. 3). Выражения RDF позволяют утверждать какую-либо информацию об элементах Semantic Web и делать выводы. Этот язык удобен также для работы с базами данных. Однако стоит помнить, что базы данных постоянно меняются, а слишком сложное описание может затруднить их быстрое обновление.

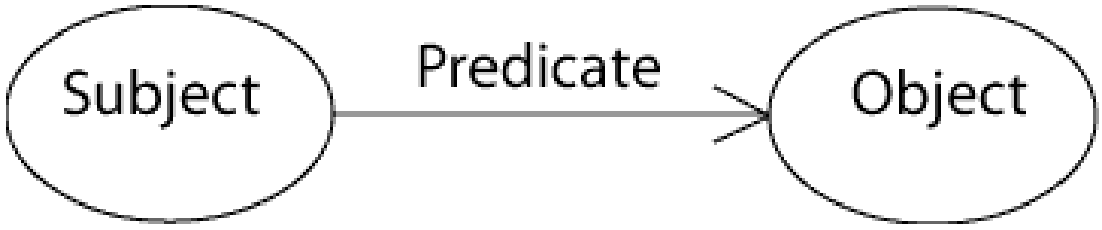

Рис. 3

Затем стал развиваться язык RDF Schema (RDFS) - язык описания словарей RDF-терминов. RDF Schema определяет классы, свойства и другие ресурсы. Таким образом, RDFS стал семантическим расширением RDF. RDF и RDFS позволяют работать с метаданными, обеспечивать компьютер семантической информацией и обрабатывать эту информацию автоматически. Однако, используя RDF, «кто угодно (т.е. любой пользователь RDF) может сказать что угодно (т.е. фиксировать произвольное утверждение) о чем угодно (т.е. о любом ресурсе Сети)». Кроме того, «RDF не запрещает делать бессмысленных утверждений или утверждений не согласующихся с другими. Следовательно, нет никакой гарантии целостности и непротиворечивости RDF-описаний» [3], поэтому при использовании информации пользователи должны сами проверять эту целостность.

Как уже упоминалось ранее, ядром Semantic Web должны были стать онтологии, однако, как отмечается в [22], отсутствие стандартизированных, единых языков онтологий затрудняло их использование внутри взаимосвязанных систем в одних и тех же областях знания. Было сложно устанавливать связи между онтологиями и соединять их. Онтологии, описывающие сходные специализированные знания, существенно

отличались в синтаксисе и семантике в зависимости от используемого онтологического языка. Это препятствовало повсеместному использованию онтологий.

Создатели Semantic Web попытались найти решение проблемы совместимости описаний, создав в 2004 году World Wide Web Consortium (W3C) для обсуждения возникших проблем. Ими был предложен универсальный стандарт для сетевого обмена онтологической информацией – Web Ontology Language (OWL). Данный язык помогает преодолеть проблему взаимодействия систем и становится основой для многих сетевых приложений. С его помощью эксперты предметной области и разработчики приложений могут создавать, модифицировать и соединять различные онтологии.

Стоит упомянуть, что сейчас существуют различные онтологические языки, однако OWL является их своеобразным универсальным объединением. Так одним из первых таких языков являлся Simple HTML Ontology Extension (SHOE), разработанный в Университете Мериленда. Этот язык был дополнен специальными тэгами для включения семантической информации в обычный код HTML. Вскоре за ним в 2000 году в Амстердамском университете был создан язык Ontology Interchange Level (OIL). Здесь уже была добавлена формальная семантика, основанная на логическом описании. Этот язык базировался на таких упомянутых ранее стандартах W3C, как XML и RDF. Одновременно с ним в США появилась программа DARPA Agent Markup Language (DAML). Она выпустила DAML-ONT, язык спецификации онтологий для реализации более сложных определений RDF классов, чем реализуемые с помощью RDFS. Позднее проект DAML соединил усилия с OIL и выпустил DAML+OIL. В DAML+OIL была усовершенствована языковая семантика, она стала яснее и успешнее взаимодействует с инструментами построения.

Вдохновленные успешным развертыванием и использованием DAML+OIL и понимая необходимость стандартизированного языка онтологий, участники W3C создали Web Ontology Language (OWL). Данный язык строится на основании RDF, который сам по себе строится на синтаксисе XML. RDF и OWL дают возможность создавать классы, свойства и отдельные экземпляры. Таким образом данный язык реализует структуру онтологий. Можно привести пример использования этих трех компонентов, предложенный в работе [22]: определим класс сущностей "People" и некоторые свойства "People", такие как "name", "birthday" и "friend".

Средствами синтаксиса RDF класс и его свойств будут описаны так:

```
<Class ID="Person"/>

<Property ID="name"/>
<Property ID="birthday"/>
<Property ID="friend"/>
```

Когда определены классы, их можно использовать при описании отдельных экземпляров. Информация “Joe Blog, born January 1, 1950, is friends with John Doe” будет представляться следующим образом:

```
<Person ID="Joe">
     <name>Joe Blog</name>
     <birthday>January 1, 1950</birthday><friend
resource="#John"/>
</Person>

<Person ID="John">
     <name>John Doe</name>
</Person>
```

Такое описание может быть подвергнуто большей степени обобщения и использоваться как общее описание класса с указанием ограничений на свойства и отношения.

Упомянутый ранее язык OWL расширяет функциональность RDF, сохраняя при этом совместимость с другими программами и являясь открытой и свободной для распространения системой. Существует несколько вариантов данного языка: OWL Lite, OWL DL, и OWL Full. Каждая разновидность характеризуется своим соотношением детальности описания и объема, сложности прикладной области:

- OWL Lite предназначен для пользователей, которым необходима лишь классификационная иерархия и некоторые простые условия согласованности сущностей.
- OWL DL (Description Logic) рассчитан на пользователей, которым необходима максимальная степень выразительных возможностей языка без потери вычислительной полноты (с гарантией получения всех возможных умозаключений, получаемых формально-логическим путем) и разрешимости (все вычисления выполняются за конечное время). Уровень OWL DL ориентирован на существующие сегодня системы описания знаний и системы логического программирования.
- OWL Full рассчитан на тех пользователей, которым необходимы максимальные выразительные возможности языка и вся свобода выбора синтаксических средств, предоставляемая в RDF, но не обязательны вычислительная полнота и разрешимость. Онтология, записанная на OWL Full, позволяет расширять значения терминов, взятых из заданных словарей. [10], [31].

OWL обладает рядом преимуществ перед RDF. К дополнительным возможностям, появившимся в этом языке, можно отнести возможность создавать локальные ограничения области распространения. Ранее у каждого свойства, отношения мог задаваться домен, но он оставался неизменным во всех областях применения. Однако порой то, как отношение применяется, зависит от конкретного класса, с которым мы работаем. Можно привести пример, упоминаемый в названной работе [22]: рассмотрим свойство «ест» для класса «Человек». Возможно, для уточнения мы захотим добавить ограничение, указав, что значение для «ест» берется из класса «Еда». Это утверждение верно для общего класса «Человек», однако оно может не распространяться на подклассы, например, подкласс «Вегетарианец». Значение «ест» для этого более узкого класса не «Еда», а ее подмножество (все, кроме мяса). Однако такая детализация была невозможна на уровне RDF, в OWL же мы можем указать локальное ограничение свойства для конкретного класса. Это уточняет описание и расширяет возможности при создании онтологии.

Кроме того, в OWL вводятся основные функции над множествами, такие как объединение, пересечение, дополнение и непересекаемость. Ценность этой возможности можно проиллюстрировать на примере: пусть в нашем распоряжении есть два подкласса "Человек", называемых "Живой Человек" и "Мертвый Человек". При использовании описания RDF не существует никаких ограничений на определение объекта, который одновременно является экземпляром класса "Живой Человек" и класса "Мертвый Человек", так как ничто не определяет эти классы взаимоисключающими. Однако эта проблема успешно решается на уровне OWL.

Еще одним преимуществом OWL является введение понятия мощности. OWL позволяет накладывать ограничения на свойство, требуя чтобы оно использовалось для любого экземпляра минимальное количество раз (минимальная мощность), максимальное количество раз (максимальная мощность) или определенное количество раз (мощность). Это означает, что можно, например, потребовать, чтобы человек-билингв говорил на двух языках.

На данном этапе OWL применяется в ряде областей, так можно привести пример системы knOWLer [13], утилиты по управлению потоками информации. Она демонстрирует, что использование онтологический умозаключений сопоставимо по результатам со стандартными системами информационного поиска. Данная система включает в себя примерно 100 миллионов утверждений. Система поддерживает сложное формирование логического вывода и работу с запросами, используя подмножество OWL. Следует упомянуть, что OWL используется также в системах, не относящихся к анализу естественного языка.

Хотелось бы отметить, что используемые ныне языки представления онтологий также рассчитаны и на моделирование нечетких знаний [24]. Например, такие расширения RDF и OWL, как Fuzzy RDF и Fuzzy OWL, позволяют совмещать стандартные модели предметных областей и нечеткую логику, применяемую на уровне задания аксиом. Известно также расширение OWL, где используются вероятностные модели знаний, например, Bayes OWL

Подводя итог, можно сказать, что OWL обеспечивает достаточно богатую семантику для описания он-лайн онтологий, которые важны для современных проектов анализа естественного языка и совместимы с web-стандартами и архитектурой. Сейчас

существуют планы дальнейшего развития OWL для усовершенствования и возможно расширения описаний.

1.7. Существующие онтологические ресурсы

Как упоминалось ранее, на данном этапе существует ряд разнообразных онтологических ресурсов. Эти онтологии, противопоставляемые по направленности, отличаются друг от друга по форме и содержанию. Опишем некоторые из онтологий, соотнося их с ранее описанными классами.

Сейчас существует ряд известных онтологий верхней зоны. Так, в качестве примера можно привести онтологическую структуры, созданной под руководством Э.Хови, которая включает 400 записей и описывает наиболее общее разбиение понятий по категориям. Известной является также Penman Upper Model, состоящая всего лишь из 300 элементов. В ней нет лексических единиц и аксиом. Ее записи представляют понятийно-грамматические классы, выступающие как связующее звено между языком и реальным миром. Между классами верхней модели и синтаксическими типами установлены связи. Эта модель, протестированная для грамматик разных языков и машинного перевода, может быть использована при построении и координации онтологий областей знаний.

Еще одним примером такого рода онтологий является DOLCE. Она была создана под руководством Н. Гуарино в Италии, в институте LADSEB (по материалам [20]). Это поверхностная модель, обладающая очень высокой степенью абстрактности. В ней содержатся примерно 500 понятий и нет лексических элементов. Описание производится с помощью языков XML или RDF, оно очень четкое и последовательное. Однако в данной онтологии малая степень иерархичности и все выглядит как верхние уровни. При конвертации этой онтологии в другие форматы могут возникать проблемы, так как пропадает часть функциональности, теряются некоторые аксиомы и ограничения.

Онтология, разработанная Дж. Совой (Sowa’s KR Ontology), также относится к категории структур верхней зоны. Эта онтология базируется на структурах, разработанных философами. Главной целью, преследуемой ее автором, являлось создание основы, в которую могут быть включены все остальные онтологии или модели их верхних зон. Структурная решетка этой онтологии представляет собой 12 понятий, являющихся всеми комбинациями трех отличительных признаков:

- первичность/вторичность/третий порядок
- материальный/абстрактный
- длительный/мгновенный (continuant/occurrent)(объект/процесс)

Составление таких решеток помогает избежать проблемы очередности использования дифференциальных признаков при делении понятий.

В онтологии Дж. Совы используются также некоторые стандартные ситуативные роли:

- детерминирующий участник (участник, определяющий направление процесса, то есть инициатор или цель);
- неотъемлемый участник (присутствует на протяжении всего процесса, но активно не контролирует происходящее);
- источник (должен присутствовать в начале процесса, но не обязан принимать участие во всем процессе);
- продукт (может появляться в конце процесса, но не обязан принимать участие во всем процессе);

К онтологиям средней зоны можно отнести Mikrokosmos, созданную в Университете Нью-Мексико. Она содержит около 6 000 записей. Онтология Принстонского университета WordNet на несколько порядков больше, она состоит из 110 000 записей, в которых в большинстве случаев отражена конкретная, нетерминологическая лексика. В ее рамках заданы различные элементы с их значениями, объединенные в группы (синсеты) по сходству значений.

CYC (**http://www.opencyc.org/**) является одной из самых крупных и разработанных онтологий. Она направлена на использование в рамках искусственного интеллекта, в ней содержится большое количество (более миллиона) аксиом. Версия ResearchCYC, преобразованная в формат RDF, не включает вторичных понятийных выражений, и таким образом, в ней пропала часть родовых понятий высшего уровня. По мнению Э.Хови [20], проблемой использования этой онтологии является запутанность ее структуры, она очень сложна для понимания.

Кроме того, на данном этапе существует множество онтологий отдельных предметных областей. Их можно разделить на несколько категорий. Так, к компьютерной области относится ряд онтологий, составленных в рамках библиотеки

онтологий Protégé, также более 700 различных онтологий были созданы в OntoSelect. Область медицины достаточно подробно описывается в UMLS, состоящей из метатезауруса (более 1 миллиона биомедицинских понятий и 5 миллионов понятийных имен), семантической сети и специального лексикона. Еще одним примером онтологии предметной области является NAICS (Североамериканская система промышленной классификации), отражающая структуру разнообразных областей (таких как технология, строительство, сельское хозяйство и другие).

Существует ряд онтологий, которые объединяют несколько более мелких и служат своего рода каркасом для встраивания новых элементов на всех уровнях. Одной из таких онтологий является Suggested Upper Model Ontology (SUMO – (http://www.ontologyportal.org/). Она и включенные в нее онтологии предметных областей представляют собой самую большую общедоступную онтологию, существующую на сегодняшний день. SUMO состоит из 20 000 элементов и 60 000 аксиом. В нее включаются сами разработки SUMO, онтология среднего уровня (MILO) и набор отраслевых онтологий (в сфере коммуникаций, транспорта, географии и многих других). Помимо этого она содержит большое количество аксиом, причем все ее элементы формально определены и это описание не зависит от конкретной системы осуществления выводов.

Онтология SENSUS также разрабатывалась как структура, в которую может быть включена дополнительная информация. Она содержит 70 000 записей и является расширением и реорганизацией WordNet; на верхнем уровне добавлены записи из Penman Upper Model и ветви WordNet переорганизованы.

Omega представляет собой среду для различных онтологий и ресурсов. Она содержит примерно 120 600 понятий и слов, в нее включены WordNet, Mikrokosmos, Penman Upper Model, а также дополнительные элементы. Это не одноязычная структура, в ней есть как английская, так и испанская лексика. Кроме того, в Omega содержится множество различных ситуативных ролей (около 13 000), взятых из ресурсов Framenet, WordNet, LCS и PropBank. Эта среда обладает обширной базой отдельных экземпляров: 1,1 миллион персоналий, 765 000 фактов, 5,7 миллионов местоположений. Omega включает огромное количество (более 28 миллионов) утверждений о понятиях, отношениях и отдельных экземплярах. Она представляет

собой ресурс с обширными возможностями построения, изменения и пополнения онтологий. Этот ресурс является библиотекой различных онтологий, позволяющей использовать объединять онтологии разных уровней для использования в какой-то прикладной задаче.

Это далеко не единственный ресурс, предоставляющий доступ к многочисленным отдельным онтологиям. Также существуют ресурсы, объединяющие онтологии, описанные с помощью какого-то языка представления, например, библиотека онтологий DAML (http://www.daml.org/ontologies/). Создатели онторедакторов (программ, упрощающих создание онтологий) предлагают своим пользователям самостоятельно формировать библиотеки онтологий, добавляя результаты своей работы прямо в специально созданный ресурс. Зачастую онтологии, содержащиеся в таких библиотеках, не отличаются общностью тематики. Однако такие ресурсы дают возможность не строить онтологии «с нуля», а использовать уже существующие наработки, находя все онтологии на одной ресурсе. Так, к таким библиотекам можно отнести библиотеку, сформированную пользователями онторедактора Protégé (ресурсы представлены на сайте - http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library). Также можно упомянуть Ontolingua (http://www.ksl.stanford.edu/software/ontolingua/), систему, разработанную в Стэнфордском университете, которая обеспечивает распределенную совместную среду для просмотра, создания, редактирования, модификации и использования онтологии.

К библиотекам онтологий относится и OntoSelect (http://olp.dfki.de/ontoselect/). Этот ресурс «отслеживает» появляющиеся онтологии в Сети, аннотирует их, предоставляя пользователям возможность в дальнейшем эффективно искать онтологии по интересующей тематике и в дальнейшем использовать их в своих целях. Также этот ресурс предоставляет возможность добавлять свою онтологию в библиотеку. На данном ресурсе можно найти ссылки на онтологии, описанные с помощью разных языков представления, так это и OWL, DAML, RDFS и другие. В OntoSelect представлены онтологии, созданные для разных языков, однако преимущественно все ресурсы разработаны для английского языка.

Заключение

Создание онтологий является перспективным направлением современных исследований по обработке информации, представляемой на естественном языке. В рамках работы освещены различные точки зрения на понятие онтологии, рассмотрены различные классификации онтологий. При создании онтологий пользователь сталкивается с рядом проблем, которые необходимо последовательно решать. Наиболее важные из этих проблем также обсуждаются в обзоре. Уже сейчас существует ряд обширных онтологий, построенных как в рамках отдельных предметных областей, так и для незамкнутых областей знания. Наиболее перспективной является автоматизация создания онтологий, однако на данном этапе еще не разработаны эффективные процедуры, применение которых позволит сократить долю ошибок. Поэтому процесс создания онтологий является столь трудоемким. Однако уже сейчас существует ряд приложений, успешно использующих онтологии в своей работе.

Библиография

1. Андреев А.М., Березкин Д.В., Рымарь В.С., Симаков К.В. Использование технологии Semantic Web в системе поиска несоответствий в текстах документов.
//URL: http://www.inteltec.ru/publish/articles/textan/rimar_RCDL2006.shtml
2. Гладун А.Я., Рогушина Ю.В. Онтологии в корпоративных системах, Часть II // Корпоративные системы №1 / 2006
//URL: http://www.management.com.ua/ims/ims116.html
3. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Курс из 16 презентаций: «Онтологии и тезаурусы». //URL: http://download.yandex.ru/class/solovyev/plan.pdf ;
(см. также: http://download.yandex.ru/class/solovyev/present1.ppt.
http://download.yandex.ru/class/solovyev/present2-1.ppt
http://download.yandex.ru/class/solovyev/present2-2.ppt
http://download.yandex.ru/class/solovyev/present3-1.ppt
http://download.yandex.ru/class/solovyev/present3-2.ppt
http://download.yandex.ru/class/solovyev/present4-1.ppt
http://download.yandex.ru/class/solovyev/present4-2.ppt
http://download.yandex.ru/class/solovyev/present5-1.ppt
http://download.yandex.ru/class/solovyev/present5-2.ppt
http://download.yandex.ru/class/solovyev/present6-1.ppt
http://download.yandex.ru/class/solovyev/present7.ppt
http://download.yandex.ru/class/solovyev/present8-1.ppt
http://download.yandex.ru/class/solovyev/present8-2.ppt
http://download.yandex.ru/class/solovyev/present9.ppt
http://download.yandex.ru/class/solovyev/present10-1.ppt
http://download.yandex.ru/class/solovyev/present10-2.ppt)
4. Добров Б.В., Лукашевич Н.В. Вторичное использование лингвистических онтологий: изменение в структуре концептуализации. //URL: http://www.rcdl2006.uniyar.ac.ru/papers/paper_78_v1.pdf

5. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска. //URL: http://fccl.ksu.ru/issue_spec/docs/oent-kgu.doc

6. Загоруйко Н.Г. и др. Система "Ontogrid" для построения онтологий //Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. конференции Диалог'2005 . М., 2005. С. 146-152.

7. Коваль С.А. Автоматическая переработка текста на базе объектно-предикатной системы // Структурная и прикладная лингвистика. Вып. 5. СПб., 1998. С. 199-207.

8. Коваль С.А. Безэкземплярные и экземплярные онтологии. Материалы XXXVI Междунар. филолог. конф. 12 марта 2007 г. в Санкт-Петербургском университете (в печати). См. также материалы к докладу на URL: http://skowal.narod.ru/research/ontology2007

9. Митрофанова О.А. Измерение семантических расстояний как проблема прикладной лингвистики // Структурная и прикладная лингвистика. Межвузовский сборник. Выпуск 7. Издательство СПбГУ, 2008.

10. Михаленко П. Язык онтологий в Web. //URL: http://www.osp.ru/os/2004/02/183921/

11. Мудрая О.В. , Бабич Б.В. , Пьяо С.С. , Рейсон П. , Уилсон Э. Разработка инструментария для семантической разметки текст // Труды международной конференции "Корпусная лингвистика–2006". Издательства СПбГУ и РХГА, 2006.

12. Bhagat R., Pantel P. and Hovy E. 2007. LEDIR: An Unsupervised Algorithm for Learning Directionality of Inference Rules. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-07). pp. 161-170. Prague, Czech Republic. //URL: http://www.patrickpantel.com/Download/Papers/2007/emnlp07.pdf

13. Ciorascu C., Ciorascu I. and Stoffel K. knOWLer - Ontological Support for Information Retrieval Systems // Proceedings of 26th Annual International ACM SIGIR Conference, Workshop on Semantic Web, Toronto, Canada, August 2003.

14. Ding Y., Fensel D., Klein M., Omelayenko B., Schulten E. The role of ontologies in eCommerce // Steffen Staab, Rudi Studer (ed.), Handbook of Ontologies, 2004.

15. Eiji Aramaki, Takeshi Imai, Masayo Kashiwagi, Masayuki Kajino, Kengo Miyo and Kazuhiko Ohe. Toward medical ontology using Natural Language Processing. //URL: http://www.m.u-tokyo.ac.jp/medinfo/ont/paper/2005-aramaki-1.pdf

16. Genesereth, M. R. and Nilsson, N. J. 1987. Logical Foundation of Artificial Intelligence. Morgan Kaufmann, Los Altos, California.

17. Gruber Th. What is an Ontology// URL: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

18. Guarino N. Understanding, Building, and Using Ontologies // URL: http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html

19. Hovy E. A Standard for Large Ontologies // URL: http://www.isi.edu/nsf/papers/hovy2.htm

20. Hovy E. Презентации Ontologies: lecture 1 , lecture 2 Issues of Content, lecture 3 Methods for Automated Ontology Building. Information Sciences Institute University of Southern California., с XX летней школы им. В.Матезиуса по лингвистике и семиотике (7-12 марта 2005 г., Карлов университет).

21. Hovy E., Knight K., Junk M. Large Resources. Ontologies (SENSUS) and Lexicons.//URL:www.isi.edu/natural-language/projects/ONTOLOGIES.html

22. Kalyanpur A. et al. OWL: Capturing Semantic Information using a Standardized Web Ontology Language. // Multilingual Computing & Technology Magazine, Vol. 15, issue 7, Nov 2004. // URL: http://www.mindswap.org/papers/MultiLing.pdf

23. Lin D., Pantel P. Concept Discovery from Text // URL: http://www.patrickpantel.com/Download/Papers/2002/coling02.pdf

24. Lim Lian-Tze and Kong Tang Enya. Building an Ontology-based Multilingual Lexicon for Word Sense Disambiguation in Machine Translation. // Proceeding of the PAPILLON-2004 Workshop on Multilingual Lexical Databases Grenoble, August

30th-September 1st, 2004. // URL: www.papillon-dictionary.org/info_media/42808971.pdf

25. Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K.J. Introduction to WordNet: an on-line lexical database. // International Journal of Lexicography 3 (4), 1990, pp. 235 - 244. // URL: ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

26. Nirenburg S., Raskin V.. Ontological Semantics. Cambridge, MA, 2004.

27. Noy N., McGuinness D. L. Ontology Development 101: A Guide to Creating Your First Ontology. // Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001. //URL:http://protege.stanford.edu/publications/ontology_development/ontology101.html

28. Pantel P. and Pennacchiotti M. 2008. Automatically Harvesting and Ontologizing Semantic Relations. In Paul Buitelaar and Philipp Cimiano (Eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text. ISBN: 978-1-58603-818-2. IOS Press. // URL: http://www.patrickpantel.com/Download/Papers/2008/olp08.pdf

29. Rosch, Eleanor. 1981. Prototype classification and logical classification: the two systems. // Ellin Scholnick (ed), New Trends in Conceptual Representation. Hillsdale, N.J.: Erlbaum, 73–85.

30. Sabou Marta. Learning WEb Service Ontologies: an Automatic Extraction Method and its evaluation. //URL: http://kmi.open.ac.uk/people/marta/papers/IOS2005.pdf

31. Smith K. Michael, Welty Chris, McGuinness L.Deborah. OWL Web Ontology Language. Guide. //URL: http://www.w3.org/TR/owl-guide/

32. Swartz A. The Semantic Web In Breadth // URL: http://logicerror.com/semanticWeb-long

33. Vacura M., Svatek V., Smrz P. A Pattern-Based Framework for Uncertainty Representation in OntologiesText, Speech and Dialogue. Proceedings of the 11th

International Conference TSD 2008, Brno, Czech Republic, September 8-12, 2008. / Eds. P. Sojka, A. Horak et al. LNAI 5246. Springer-Verlag, 2008.

34. Van Heijst, G., Schreiber, A. T., and Wielinga, B. J. 1996. Using Explicit Ontologies in KBS Development. //URL: http://ksi.cpsc.ucalgary.ca/KAW/KAW96/borst/node16.html

35. Wielinga, B. J. and Schreiber, A. T. Reusable and sharable knowledge bases: a European perspective // Proceedings of First International Conference on Building and Sharing of Very Large-Scaled Knowledge Bases. Tokyo, Japan Information Processing Development Center. 1993.

36. Wierzbicka A. Semantic primitives. Frankfurt, 1972.