ПОДХОД К ПОСТРОЕНИЮ ПРЕДМЕТНОЙ ОНТОЛОГИИ ДЛЯ ПОРТАЛА ЗНАНИЙ ПО КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ[1]

APPROACH TO BUILDING SUBJECT ONTOLOGY FOR KNOWLEDGE PORTAL COMPUTATIONAL LINGUISTICS

Ю.А. Загорулько *(zagor@iis.nsk.su)*,

О.И. Боровикова *(olesya@iis.nsk.su)*,

И.С. Кононенко *(irina_k@cn.ru)*,

Е.А. Сидорова *(lena@iis.nsk.su)*
Институт систем информатики имени А.П.Ершова СО РАН,
Российский НИИ Искусственного Интеллекта, Новосибирск

Рассматривается подход к построению онтологии для портала знаний по компьютерной лингвистике. Онтология портала включает предметную онтологию, а также онтологии научной деятельности и научного знания. Предметная онтология строится на основе онтологии научного знания и включает иерархии разделов компьютерной лингвистики, научных результатов, объектов и методов исследования.

Введение

В настоящее время в сети Интернет представлен большой объем знаний и информационных ресурсов по компьютерной лингвистике [1]. Однако эти ресурсы не достаточно систематизированы, т.е. практически случайным образом распределены по каталогам и электронным архивам или размещены на отдельных сайтах как гуманитарной, так и технической направленности, что значительно затрудняет их поиск и использование. В то же время в связи с нарастающими потребностями в естественно-языковых, в том числе речевых, интерфейсах и средствах автоматической обработки документов, возникает необходимость в эффективном доступе к публикациям, описывающим методы и подходы к обработке естественно-языковых текстов, разного рода словарям, программным компонентам и алгоритмам, реализующим ту или иную задачу или этап обработки (анализа/синтеза) текста или речи.

Чтобы удовлетворить описанную выше потребность, предлагается построить специализированный Интернет-портал знаний [2], который обеспечит систематизацию знаний и информационных ресурсов по компьютерной лингвистике, их интеграцию в единое информационное пространство, а также содержательный доступ (поиск информации в терминах предметной области портала) и удобную навигацию по нему.

Пользователями такого портала могли бы стать как научные работники, преподаватели и студенты, исследующие, преподающие и изучающие эту дисциплину, так и специалисты, разрабатывающие программные системы, предназначенные для обработки текстов, анализа и синтеза речи.

Информационную основу портала составляет онтология, подразделяющаяся на универсальную онтологию науки, служащую для представления понятий, необходимых для описания научной деятельности и научного знания в целом, и онтологию предметной области, представляющую компьютерную лингвистику как научную дисциплину. Последняя онтология определяет систематизацию знаний и информационных ресурсов, а следовательно удобство доступа к ним. Поэтому именно от нее, в конечном счете, зависит полезность портала для пользователей, описанных выше типов. Разработке этой онтологии и посвящена данная статья.

Требования к онтологии портала знаний

Как было сказано выше в качестве информационной (концептуальной) основы портала была выбрана онтология. Опишем, что мы понимаем под онтологией.

Онтология – это шестерка вида:

<C, A, T, D, R, F>, где

C – множество классов, описывающих понятия некоторой предметной или проблемной области;

A – множество атрибутов, описывающих свойства понятий;

T – множество типов значений атрибутов;

D – множество доменов;

[1] Работа выполняется при финансовой поддержке РФФИ (проект № 04-01-00884)

R – множество отношений, заданных на классах (понятиях);

F – множество ограничений на значения атрибутов.

Вводя таким образом формальные описания понятий (в виде классов объектов) и отношений между ними, онтология задает структуры для представления реальных объектов и событий, существующих в некоторой предметной или проблемной области, и обеспечивает их взаимосвязи.

В процессе разработки онтологии выделяются и формально описываются классы понятий, связанные в иерархию с помощью отношения наследования. Различные свойства каждого понятия описываются с помощью атрибутов понятий и ограничений, наложенных на область их значений. Механизм наследования определен таким образом, что наследующему понятию от родительского понятия передаются не только все атрибуты, но и отношения.

Процесс разработки онтологии (согласно методологиям, представленным в работах [3–6]) также должен включать определение возможной области применения онтологии и задач, которые могут быть решены с ее помощью. Для того, чтобы онтология удовлетворяла целям портала, она должна обеспечивать:

простую настройку портала на выбранную область знаний;

интеграцию знаний и информационных ресурсов в единое информационное пространство;

содержательный доступ и удобную навигацию по всему информационному пространству портала.

Для упрощения настройки портала на выбранную область знаний в онтологии портала необходимо выделить структуры, независимые от предметной области (ПО) портала.

Чтобы обеспечить интеграцию знаний и информационных ресурсов в единое информационное пространство, онтология должна не только представлять формальное описание системы понятий проблемной и предметной областей портала, но на ее основе также должны описываться типы информационных ресурсов и их связи с другими понятиями онтологии.

Онтология должна обеспечить такое представление свойств понятий и отношений между ними, на основе которого можно было бы автоматически строить внутренние хранилища данных портала, включающие экземпляры классов понятий и отношений, определенных в онтологии, осуществлять навигацию по информационному пространству портала и организовывать содержательный поиск.

Для того, чтобы онтология могла играть такую важную роль, она должна быть хорошо структурирована и адекватно отражать проблемную и предметную область портала. В связи с этим в онтологии портала выделяются предметно-независимые (базовые) онтологии и онтология предметной области.

В качестве базовых выбраны разработанные ранее онтологии научной деятельности и научного знания [7], которые не зависят от предметной области портала. Для описания предметной области портала служит онтология ПО или предметная онтология.

Предметно-независимые онтологии

Рассмотрим подробнее онтологии научной деятельности и научного знания (Рис.1).

Онтология научной деятельности включает базовые классы понятий, относящиеся к организации научной и исследовательской деятельности, такие как Персона, Организация, Событие, Публикация, Информационный ресурс.

Класс *Персона* служит для представления субъектов научной деятельности: исследователей, сотрудников и членов организаций и т.п.

Класс *Организация* включает понятия, которые описывают различные организации, научные сообщества, институты, исследовательские группы и другие объединения.

В класс *Событие* входят понятия, описывающие научно-организационную или научно-исследовательскую деятельность. В этом классе выделяются подклассы Научное мероприятие и Деятельность. Первый подкласс служит для описания таких научных мероприятий как семинары, конференции, выставки и т.п. Понятия класса Деятельность являются связующим звеном между методом и объектом исследования и полученным научным результатом. Класс описывает такие понятия, как Проект, Программа исследований и т.п.

Класс *Публикация* служит для описания различных типов публикаций и материалов, представленных в печатном или электронном формате (монографии, статьи, отчеты, труды конференций, периодические издания, фото- и видео-материалы и др.).

Класс *Информационный ресурс* служит для описания различных информационных ресурсов, представленных в сети Интернет.

Онтология научного знания, по своей сути, является метаонтологией. Она содержит метапонятия и отношения, задающие структуры для описания рассматриваемой предметной области, такие как Раздел науки, Метод исследования, Объект исследования, Научный результат, позволяющие выделить в данной науке значимые разделы и подразделы, задать типизацию методов и объектов исследования, описать результаты научной деятельности.

Понятия онтологии научного знания связаны между собой и понятиями онтологии научной деятельности следующими ассоциативными отношениями:

«научное направление» – позволяет связывать события, публикации, организации, исследователей, информационные ресурсы с разделами науки;

«описывает» – задает связь публикации с научным результатом, объектом или методом исследования;

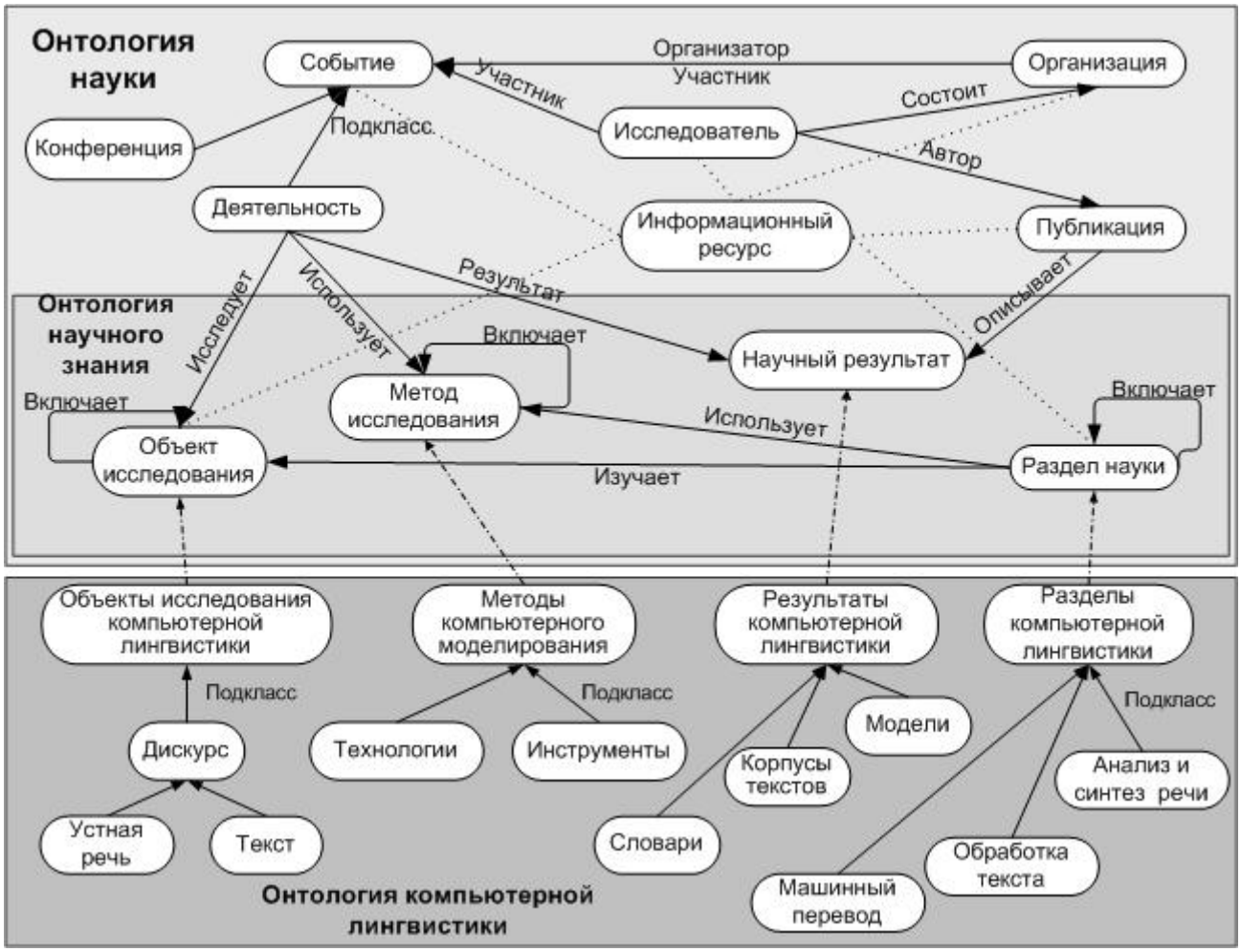

Рис. 1 Онтология портала

«использует» – связывает метод исследования с деятельностью или разделом науки;

«исследует» и *«изучает»* – сопоставляет соответственно деятельность и раздел науки с объектом исследования;

«результат деятельности» – связывает научный результат с деятельностью;

«ресурс» – связывает информационный ресурс с методами и объектами исследования.

Следует отметить, что выбор описанных выше ассоциативных отношений осуществлялся не только исходя из полноты представления проблемной и предметной областей портала, но и с учетом удобства навигации по его информационному пространству и поиска информации.

Предметная онтология строится на основе базовых онтологий и отражает общие знания о предметной области, такие как иерархия классов понятий, семантические отношения на этих классах. Более подробно структура и процесс построения предметной онтологии рассматривается в следующей главе.

Построение онтологии предметной области

Онтология предметной области рассматриваемого портала знаний описывает компьютерную лингвистику в целом как раздел науки. Она строится для организации эффективного доступа к знаниям и информационным ресурсам по данной тематике и, следовательно, должна соответствовать требованиям, приведенным выше.

Важным моментом при разработке онтологии предметной области является построение иерархий понятий. При выборе подходящей иерархии необходимо учитывать, не только то, насколько она полно описывает предметную область и связанные с ней информационные ресурсы, но и насколько удобно пользователю с ее помощью осуществлять навигацию по информационному пространству портала и вести содержательный поиск.

Таким образом, при построении иерархии нужно учитывать, что в дальнейшем на ее основе: автоматически строится схема базы данных портала (генерация структуры БД и ее ограничений целостности должна выполняться по всем элементам онтологии); создаются формы для заполнения БД портала данными (экземплярами классов понятий и отношений онтологии); определяется схема навигации по информационному пространству портала (по отношениям онтологии); генерируются формы поисковых запросов (по классам и отношениям онтологии).

Предложенная нами онтология компьютерной лингвистики включает четыре базовых иерархии: иерархия разделов компьютерной лингвистики, в основе которой лежат классификации программных технологий и основных теоретических направлений компьютерной лингвистики, иерархия объектов, иерархия методов исследования и иерархия научных результатов.

Иерархия разделов науки определяет значимые разделы и подразделы компьютерной лингвистики и служит для описания выделенного в объекте исследования предмета интереса, определяющего действия, которые можно совершать с объектом в зависимости от целей исследователя. К разделам компьютерной лингвистики относятся, например, такие направления как Машинный перевод, Обработка текста, Анализ и синтез речи, и др. Эти общие направления также подразделяются на более частные. Например, Машинный перевод включает Автоматический и Автоматизированный машинный перевод.

Иерархия методов исследования служит для систематизированного описания инструментов исследования, применяемых в компьютерной лингвистике. Здесь выделяются такие подразделы, как Методы, Технологии, Системы.

Иерархия объектов исследования задает типизацию объектов исследования и структуры для их описания. В качестве базового объекта исследования рассматривается Дискурс как форма существования и использования языка (языковых единиц различных уровней в их системной взаимосвязи). В частности, учитываются фонетические, морфологические, синтаксические и другие языковые явления, а также такие формы Дискурса, как Текст и Устная речь.

Иерархия научных результатов служит для типизации и описания результатов научной деятельности. Она включает такие типы результатов, как Модели, Словари, Формальные описания, Корпусы текстов.

Все иерархии онтологии компьютерной лингвистики связаны между собой посредством ассоциативных отношений, семантика которых была определена при описании базовых онтологий.

Заключение

В данной работе предложен только подход к построению предметной онтологии портала знаний по компьютерной лингвистике и показаны ее базовые структуры и элементы. Детальная разработка этой онтологии является ближайшей целью авторов. Предполагается также работа по сбору и описанию информационных ресурсов, связыванию их с понятиями онтологии портала.

При построении онтологии компьютерной лингвистики будут использоваться научные отчеты и обзоры [8], материалы сайтов, Интернет-каталогов, электронных журналов и энциклопедий [9–15], труды конференций по компьютерной лингвистике (в том числе конференции «Диалог» [16]), а также знания и опыт авторов доклада и их коллег.

Список литературы:

1. *Захаров В.П., Булдакова Е.В. // Международный форум по информатике. 2001.Т. 26, № 1, С.30-36.*
2. *Боровикова О.И., Загорулько Ю.А. Организация порталов знаний на основе онтологий. // Труды международного семинара Диалог'2002 "Компьютерная лингвистика и интеллектуальные технологии". Протвино, 2002. Т.2, С.76-82.*
3. *Uschold M., Gruninger M. Ontologies: Principles, Methods and Applications // Knowledge Engineering Review11(2), 1996.*
4. *Gruninger M., Fox M.S. Methodology for the Design and Evaluation of Ontologies // Proceedings of IJCAI 1995 Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.*
5. *Fernandez-Lopez M., Gomez-Perez A., Pazos Sierra J. Building a Chemical Ontology Using Methontology and the Ontology Design Environment // IEEE Intelligent Systems, 1999. 14(1), P.37-46.*
6. *Staab S., Schunurr H-P., Studer R., Sure Y. Knowledge processes and ontologies // IEEE Intelligent Systems, Special Issue on Knowledge Management, 2001. 16(1), P.26-34.*
7. *Zagorulko Yu., Borovikova O., Bulgakov S., Sidorova E. Ontology-based approach to development of adjustable knowledge internet portal for support of research activity // Bull. of NCC. Ser.: Comput. Sci. 2005. Is. 23, P.45–56.*
8. *Survey of the State of the Art in Human Language Technology editors Cole R.A., Mariani J., Uszkoriet H., Zaenen A., Zue V. // Standford University, Stanford, CA, Cambridge University Press, 1996.*
9. *http://www.krugosvet.ru/ Энциклопедия «Кругосвет».*
10. *http://www.i-u.ru/biblio/dict.aspx/ Словари на сайте Русского Гуманитарного Интернет-Университета.*
11. *http://www.philol.msu.ru/rus/ Филологический факультет МГУ.*
12. *http://speech.bme.ogi.edu/ CSLU - Research Center for Spoken Language Understanding.*
13. *http://www.aboutai.net/ Раздел Natural Language Understanding на сайте AboutAI.net.*
14. *http://linguistlist.org/ The LINGUIST List.*
15. *http://acl.ldc.upenn.edu/ A Digital Archive of Research Papers in Computational Linguistics.*
16. *http://www.dialog-21.ru/ Материалы конференции ДИАЛОГ.*