ПОРТАЛ ЗНАНИЙ ПО КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ: СОДЕРЖАТЕЛЬНЫЙ ДОСТУП К ЛИНГВИСТИЧЕСКИМ ИНФОРМАЦИОННЫМ РЕСУРСАМ[1]

Загорулько Ю.А. *(*zagor@iis.nsk.su*)*, ***Боровикова О.И.*** *(*olesya@iis.nsk.su*)*, ***Загорулько Г.Б.*** *(*gal@iis.nsk.su*)*

Институт систем информатики имени А.П.Ершова СО РАН, Новосибирск, Россия

Рассматривается Интернет-портал знаний, обеспечивающий систематизацию знаний и информационных ресурсов по компьютерной лингвистике, их интеграцию в единое информационное пространство, а также содержательный доступ к ним (поиск информации в терминах предметной области портала и управляемую знаниями навигацию).

Введение

В связи с постоянно растущими потребностями в средствах автоматической обработки документов и естественно-языковых, в том числе речевых, интерфейсах, возникает необходимость в эффективном доступе не только к публикациям, описывающим методы и подходы к обработке текстов, но и разного рода словарям, программным компонентам и алгоритмам, реализующим различные задачи обработки текста или речи. И, хотя в настоящее время в сети Интернет представлен большой объем знаний и информационных ресурсов по этой тематике, доступ к таким ресурсам значительно затруднен, так как они лишь частично систематизированы и при этом рассредоточены по различным Интернет-сайтам, каталогам и электронным архивам.

Для решения этой проблемы существует несколько подходов. В рамках одного из них создаются различные Интернет-ресурсы (форумы, рассылки, комьюнити-порталы), выполняющие информационную поддержку разнообразных тематических сообществ. Самым известным ресурсом такого рода, имеющим отношение к компьютерной лингвистике, является англоязычный каталог LINGUIST List (http://linguistlist.org/), созданный для общения и обмена знаниями между лингвистами и содержащий информацию о публикациях, персоналиях, научных учреждениях и других организациях лингвистического направления, грантах, конкурсах, проектах, фондах и источниках финансирования, конференциях и семинарах лингвистической тематики. LINGUIST List предоставляет возможность поиска ресурсов по таким параметрам, как страна, язык, раздел лингвистики.

К российским аналогам LINGUIST List можно отнести научно-образовательный портал "Лингвистика в России: ресурсы для исследователей" (http://uisrussia.msu.ru/linguist/index.jsp) и сайт “Российская лингвистика (RUSLING)” (http://rusling.narod.ru), создаваемый в Отделении лингвистических исследований ВИНИТИ РАН. Портал "Лингвистика в России” содержит иерархически организованный каталог ссылок на наиболее значимые лингвистические ресурсы и позволяет осуществлять навигацию по разделам портала с помощью иерархических связей внутри этих разделов и по ссылкам на связанные с ними области (разделы). Тематические категории этого портала представлены разделами по компьютерной, теоретической и прикладной лингвистике и их приложениям (смежным областям), а также разделами, посвященными русскому языку, языкам мира и народов РФ. Портал “Российская лингвистика” предлагает лингвистам «информационную карту» для поиска информации об организациях, научных исследованиях и публикациях, лингвистических ресурсах и персоналиях. Он содержит обширный каталог ссылок на словари и корпуса текстов для различных языков (в том

[1] Работа выполняется при финансовой поддержке РГНФ (проект № 07-04-12149)

числе славянских), а также сведения о российских лингвистах, предоставляя возможность их поиска не только по алфавиту, но и по области и объекту (языку) исследования.

Информационное наполнение порталов такого типа в значительной мере зависит от способа сбора информации (его автоматизированности) и личного вклада и активности каждого участника сообщества.

Другой подход направлен на представление лингвистических ресурсов непосредственно для работы с лингвистическими данными. К таким проектам относятся работы по переводу текстов в цифровые форматы, созданию средств их хранения и обработки, построению лингвистических онтологий и web-интерфейсов для описания и наполнения ресурсов лингвистическими данными. Среди таких проектов можно отметить проект E-MELD (http://emeld.org), в рамках которого создается лингвистическая онтология GOLD (General Ontology for Linguistic Description), представляющая общеязыковые знания в виде иерархических структур.

Как правило, проекты, разрабатываемые в рамках описанных выше подходов, направлены на описание и сохранение общеязыковой лингвистической информации, а не для интеграции ресурсов по компьютерной лингвистике и обеспечения к ним содержательного доступа широкому кругу пользователей.

Для решения этой проблемы нами разрабатывается специализированный Интернет-портал знаний – портал знаний по компьютерной лингвистике. Как информационный ресурс такой портал знаний обеспечивает следующие возможности:

- представление научной дисциплины «компьютерная лингвистика» (используемых в ней терминов и понятий, тематических разделов, объектов и методов исследования, научных результатов и т.п.) и участников научной деятельности в рамках этой дисциплины (персоналий, групп, сообществ и других организаций, включенных в процесс исследования);
- интеграцию доступных информационных ресурсов по компьютерной лингвистике в единое информационное пространство;
- содержательный доступ к систематизированным знаниям и данным, относящимся к компьютерной лингвистике, т.е. возможность поиска и получения информации в терминах предметной области портала, а также удобную навигацию по всему информационному пространству портала, базирующуюся на модели предметной области;
- персонификацию пользовательского интерфейса (способа и степени подробности предоставления информации, поиска и навигации по порталу);
- информационную поддержку пользователей, т.е. анонсирование разного рода событий и мероприятий, касающихся данной дисциплины.

1. Информационная модель портала

Информационная модель портала должна обеспечивать унифицированное представление и хранение знаний и информационных ресурсов по компьютерной лингвистике, а также содержательный доступ к ним: поиск информации в терминах предметной области портала и удобную навигацию по его информационному пространству. Поэтому в качестве основы такой модели выбрана онтология [1], содержащая наряду с традиционным описанием проблемной и предметной областей соотнесенное с ним описание соответствующих сетевых ресурсов [2].

С содержательной точки зрения, онтология портала служит для представления понятий, необходимых для описания как научной деятельности и научного знания в целом, так и конкретной научной дисциплины, в частности. В связи с этим онтология портала включает универсальные онтологии научной деятельности и научного знания [3], а также онтологию предметной области.

Первые две из перечисленных онтологий не зависят от предметной области (ПО) и могут использоваться практически в любом портале знаний, независимо от его тематики. В связи с этим эти онтологии выделе в качестве базовых (Рис.1). Рассмотрим их подробнее.

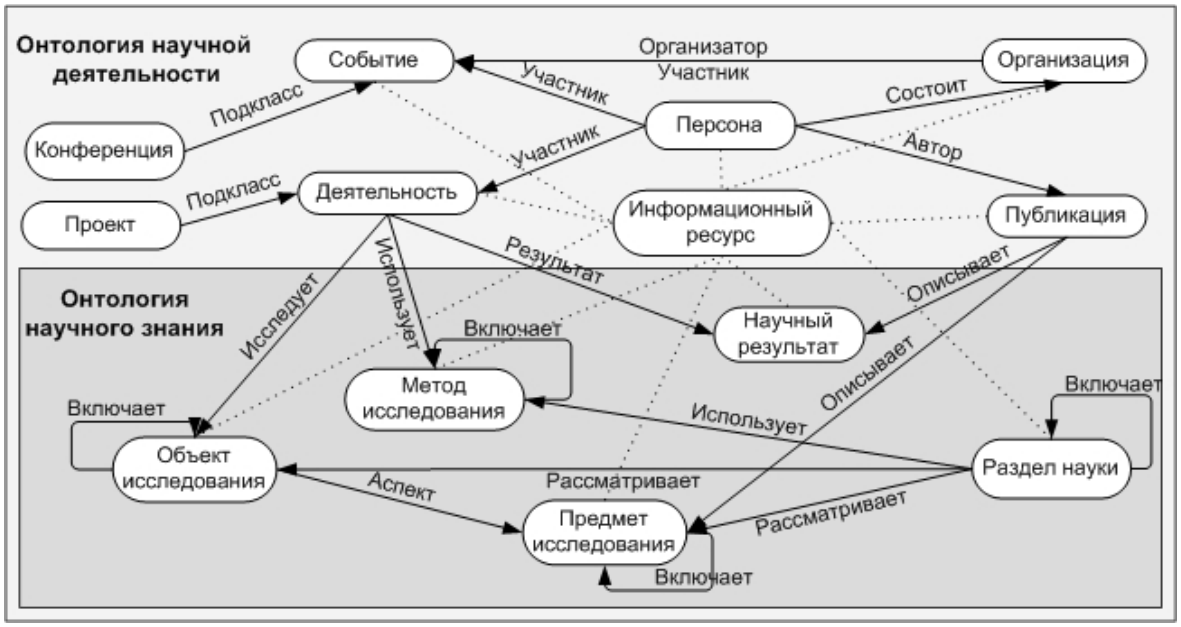


Рис. 1 Базовые онтологии портала

Онтология научной деятельности является онтологией верхнего уровня и включает базовые понятия, относящиеся к организации научно-исследовательской деятельности, такие как *Персона*, *Организация*, *Событие*, *Деятельность*, *Публикация*, используемые для описания участников научной деятельности, мероприятий, научных программ и проектов, различного типа публикаций. В эту онтологию также включено понятие *Информационный ресурс*, которое служит для описания 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*, *Предмет исследования*, *Объект исследования*, *Метод исследования*, *Научный результат*, позволяющие выделить в данной науке значимые разделы и подразделы, задать типизацию предметов, объектов и методов исследования, описать результаты научной деятельности.

Понятия базовых онтологий связаны между собой ассоциативными отношениями (см. Рис.1), выбор которых осуществлялся не столько исходя из полноты представления проблемной и предметной областей портала, сколько исходя из удобства навигации по его информационному пространству и поиска информации. Свойства каждого понятия описываются с помощью атрибутов и ограничений, наложенных на область их значений.

Так как портал предназначен для организации содержательного доступа к лингвистическим ресурсам, то в качестве онтологии предметной области он включает онтологию компьютерной лингвистики (КЛ). Понятия этой онтологии являются реализациями метапонятий онтологии научного знания и организованы в 5 иерархий «общее-частное»: *Иерархия Объектов исследования*, *Иерархия Предметов исследования*, *Иерархия Методов исследования*, *Иерархия Разделов науки*, *Иерархия Научных результатов* (см. Рис.2). Все эти иерархии связаны между собой посредством ассоциативных отношений, часть которых наследуется из базовых онтологий, а часть отражает специфику данной предметной области.

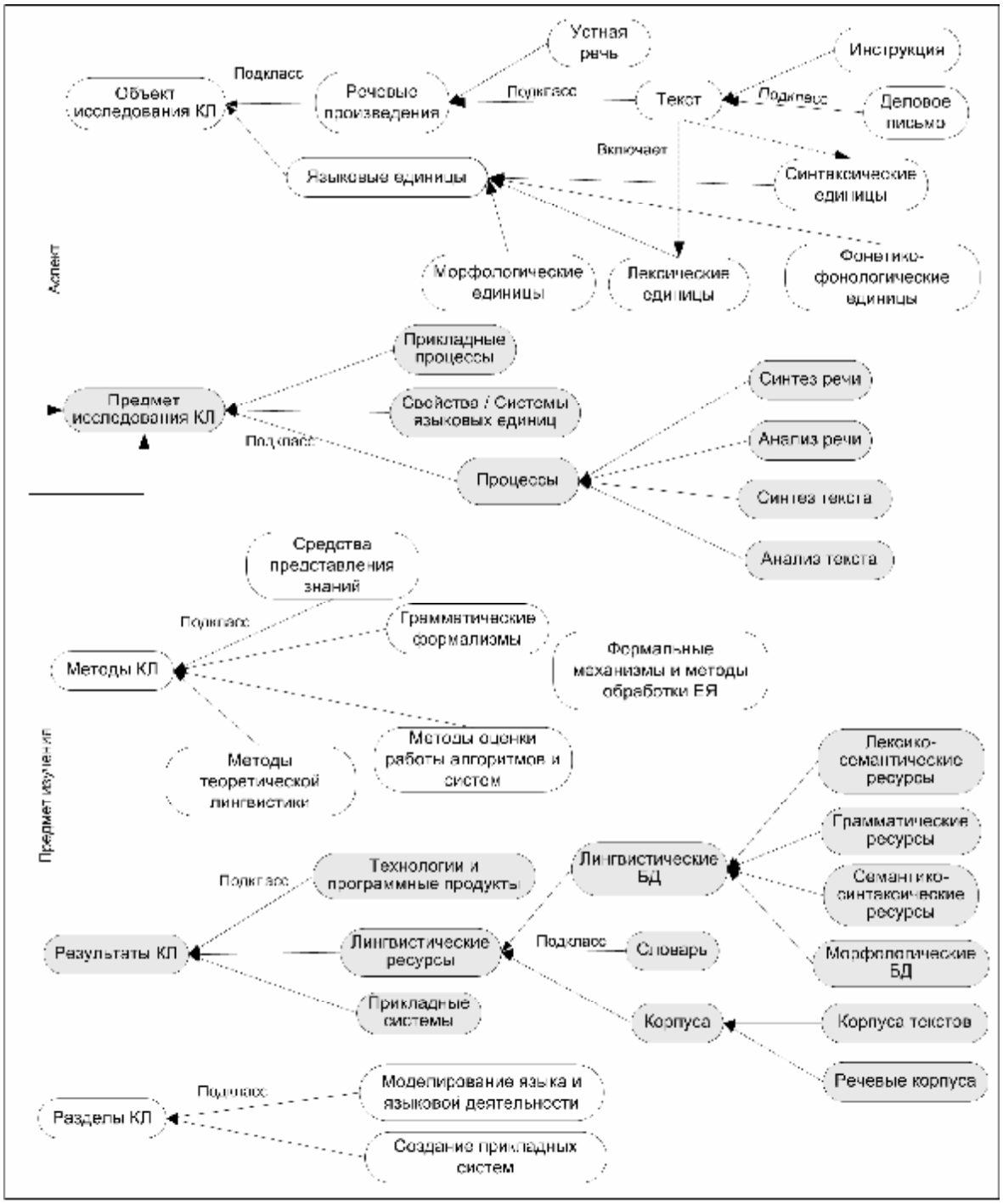


Рис. 2 Ядро онтологии компьютерной лингвистики

Таким образом, вводя формальные описания понятий проблемной и предметной области в виде понятий и отношений между ними, онтология портала задает структуры для представления реальных объектов и связей между ними.

В соответствии с принятой моделью данные на портале представлены в виде множества разнотипных информационных объектов и связей между ними. *Информационный объект* (ИО) – это структурированная совокупность данных, представляющая описание некоторого объекта выбранной области знаний или релевантного ей информационного ресурса. Каждый ИО соответствует некоторому понятию онтологии (является экземпляром этого понятия) и имеет заданную им структуру. Между конкретными информационными объектами могут существовать связи,

семантика которых определяется отношениями, заданными между соответствующими понятиями онтологии.

2. Информационное содержание портала

Информационное содержание (контент) портала включает как знания общего характера (представлены в онтологии), так и конкретные знания о реальных объектах и информационных ресурсах, систематизированные в соответствии с онтологиями портала.

Так как портал посвящен компьютерной лингвистике, в его контенте, в первую очередь, представлены знания об основных разделах компьютерной лингвистики, о ее предметах и объектах исследования, используемых в ней моделях и методах, разработанных прикладных и инструментальных системах, алгоритмах и лингвистических ресурсах. Кроме этого, пользователи портала могут получить представление не только о компьютерной лингвистике как научной дисциплине, но и найти информацию о выполняемой в этой области научной деятельности. В первую очередь, это информация об ученых, исследовательских группах и организациях и их деятельности.

В деятельности организаций и исследователей особое место занимают научные и коммерческие проекты, в рамках которых большей частью и создаются лингвистические знания и ресурсы. Результаты этой деятельности находят отражение в публикациях - монографиях, статьях, материалах конференций и семинаров, отчетах и других текстовых ресурсах, доступ к которым обеспечивается порталом.

Таким образом, ресурсы компьютерной лингвистики представлены непосредственно результатами деятельности организаций и отдельных исследователей, полученных в ходе выполнения научных и коммерческих проектов. К таким ресурсам относятся как технологии, программные продукты, прикладные системы, так и лингвистические ресурсы: словари, корпуса и лингвистические БД. Для организации более эффективного доступа к таким ресурсам в контенте представлена информация о различных аспектах их разработки: организациях, персонах и проектах, с которыми связано их появление, а также о таких содержательных характеристиках ресурсов, как отнесенность к разделу науки, объекту или предмету исследования, методам исследования. Эта информация связывает ресурсы с остальными данными и знаниями, представленными в контенте портала, что позволяет пользователю выделить группы ресурсов, созданные, например, в ходе осуществления некоторой исследовательской деятельности (гранта, проекта, конкурса) или с использованием определенного класса методов исследования.

Важным компонентом информационного контента портала является описание Интернет-ресурсов. К таким ресурсам относятся сайты организаций, конференций, проектов, порталы и каталоги, а также отдельные страницы с материалами графического, мультимедийного или текстового типа. Как было сказано выше, каждый Интернет-ресурс, представленный на портале, соответствует такому понятию онтологии, как Информационный ресурс. Описание каждого ресурса включает экземпляр данного понятия (информационный объект) и набор экземпляров отношений, связывающих данный ИО с другими информационными объектами, представляющими организации, персоны, публикации, события, разделы науки и т.д.

3. Настройка портала и управление его контентом

Настройка портала на предметную область и управление его информационным контентом осуществляются с помощью специализированных редакторов (редактора онтологии и редактора данных), реализованных как web-приложения и доступных зарегистрированным пользователям через Интернет, а также коллекционера онтологической информации о ресурсах.

С помощью редактора онтологии можно создавать, модифицировать и удалять любые элементы онтологии: понятия, отношения, домены, задавать и модифицировать иерархии понятий.

Для более удобного представления информации пользователю портала в редактор онтологий также включены средства настройки визуализации знаний и данных. Эти средства позволяют для каждого понятия онтологии задать шаблон визуализации объектов – экземпляров этого понятия и шаблон визуализации ссылок на них.

Редактор данных, функционирование которого основано на онтологии портала, позволяет создавать, редактировать и удалять информационные объекты, а также связывать их с введенными ранее объектами и понятиями.

Коллекционер онтологической информации о ресурсах предназначен для автоматизации сбора релевантных Интернет-ресурсов [4]. Он включает модуль сбора информации и модуль автоматического индексирования и классификации.

Модуль сбора информации обеспечивает поиск текстовых ресурсов или документов по ключевым словам, характеризующим область знаний портала, на сайтах и Интернет-страницах, ссылки на которые хранятся в специальной базе данных. Эта база данных может пополняться вручную (экспертом) или автоматически (за счет ссылок, обнаруженных в скачанных документах).

Модуль автоматического индексирования и классификации, используя онтологию и предметный словарь, строит содержательный индекс (семантическую аннотацию) для каждого документа и определяет раздел науки, к которому он относится. Затем эти данные представляются в информационном пространстве портала в виде информационных объектов и их связей и могут быть использованы при поиске информации и навигации.

4. Обеспечение доступа к ресурсам по компьютерной лингвистике

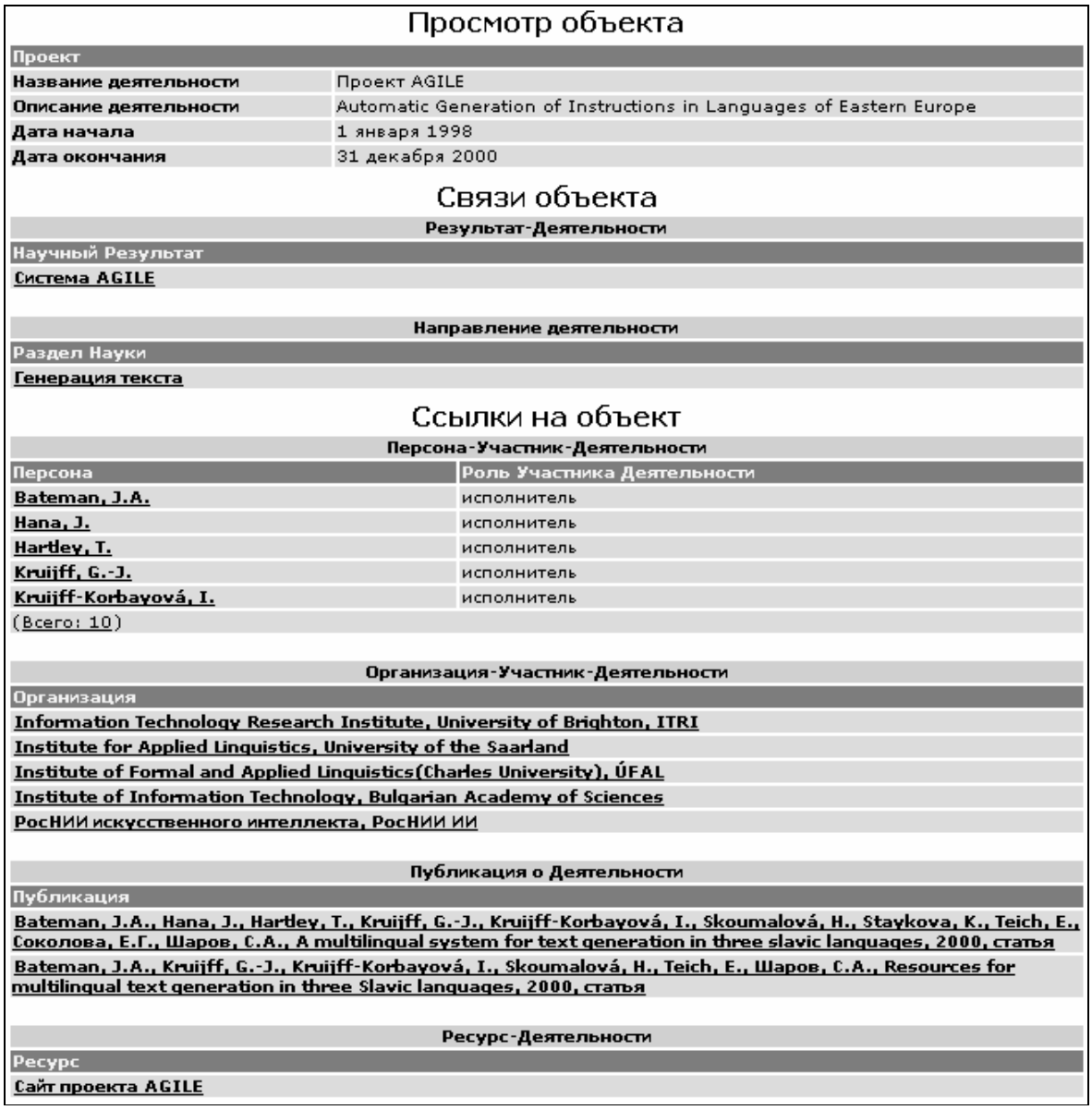
Основное назначение рассматриваемого портала знаний – обеспечить содержательный доступ к систематизированным знаниям и информационным ресурсам по компьютерной лингвистике. Доступ к знаниям и данным портала осуществляется путем навигации по дереву понятий онтологии и информационному пространству портала, а также через развитые средства содержательного поиска (с использованием понятий и отношений онтологии).

4.1. Навигация по информационному пространству портала

Для конечного пользователя данные на портале представлены в виде множества связанных информационных объектов. При навигации по порталу обеспечивается возможность выбора ИО, относящихся к интересующему нас понятию, просмотра и фильтрации списков выбранных ИО, навигации по конкретным ИО, а также просмотра описания выбранного нами информационного ресурса.

Список ИО отображается в виде страницы, содержащей набор ссылок на эти объекты. Для больших списков формируется составная страница, включающая список страниц с элементами навигации по этому списку.

Вся информация о конкретном объекте и его связях отображается в виде HTML-страницы (Рис.3), формат и наполнение которой зависят от свойств понятия, экземпляром которого является данный объект, и заданного для него шаблона визуализации. При этом объекты, связанные с данным объектом, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.



Просмотр объекта

Проект	
Название деятельности	Проект AGILE
Описание деятельности	Automatic Generation of Instructions in Languages of Eastern Europe
Дата начала	1 января 1998
Дата окончания	31 декабря 2000

Связи объекта

Результат-Деятельности

Научный Результат
Система AGILE

Направление деятельности

Раздел Науки
Генерация текста

Ссылки на объект

Персона-Участник-Деятельности

Персона	Роль Участника Деятельности
Bateman, J.A.	исполнитель
Hana, J.	исполнитель
Hartley, T.	исполнитель
Kruijff, G.-J.	исполнитель
Kruijff-Korbayová, I.	исполнитель

(Всего: 10)

Организация-Участник-Деятельности

Организация
Information Technology Research Institute, University of Brighton, ITRI
Institute for Applied Linguistics, University of the Saarland
Institute of Formal and Applied Linguistics(Charles University), ÚFAL
Institute of Information Technology, Bulgarian Academy of Sciences
РосНИИ искусственного интеллекта, РосНИИ ИИ

Публикация о Деятельности

Публикация
Bateman, J.A., Hana, J., Hartley, T., Kruijff, G.-J., Kruijff-Korbayová, I., Skoumalová, H., Staykova, K., Teich, E., Соколова, Е.Г., Шаров, С.А., A multilingual system for text generation in three slavic languages, 2000, статья
Bateman, J.A., Kruijff, G.-J., Kruijff-Korbayová, I., Skoumalová, H., Teich, E., Шаров, С.А., Resources for multilingual text generation in three Slavic languages, 2000, статья

Ресурс-Деятельности

Ресурс
Сайт проекта AGILE

Рис. 3 Представление информационного объекта и его связей

Таким образом, навигация по данным портала представляет собой процесс перехода от одних информационных объектов к другим по заданным между ними связям.

Например, при просмотре информации о конкретном проекте (см. Рис.3) мы можем видеть значения его атрибутов и его связи с другими объектами. Используя представленные связи в качестве элементов навигации, можно перейти к просмотру подробной информации как по прямым связям (об объекте исследования, об используемых методах и научных результатах, полученных в ходе выполнения проекта), так и по обратным (об участниках проекта, публикациях о проекте, информационном ресурсе, описывающем данный проект).

При переходе по конкретной связи любого информационного объекта мы можем получить достаточно большой список объектов (например, список людей, работающих в некоторой организации). В связи с этим был введен механизм фильтрации списков информационных объектов. Фильтрация есть способ выборки подмножества ИО из списка путем задания условий, которые определяют допустимые значения атрибутов ИО и требования к существованию связей с другими информационными объектами. Этот метод позволяет, например, отфильтровать множество публикаций как по дате публикации

(условия на атрибут), так и по описываемому научному результату или объекту исследования (условия на связанный объект).

4.2. Поиск в терминах предметной области

При поиске информации пользователю предоставляется возможность задания запроса в терминах предметной области портала. При этом пользователь должен выбрать понятие, к которому относятся искомые информационные объекты, и определить ограничения, которым должны удовлетворять атрибуты выбранного понятия и его связи с другими понятиями.

Ограничения на отдельные атрибуты интерпретируются как конъюнкция условий. Допустимые ограничения для атрибута зависят от типа его значений. Так, например, для атрибутов типа «integer» и «date» задается точное значение или допустимый интервал значений.

Пользователю также предоставляется возможность задать условия на значения атрибутов объектов, связанных с искомым объектом. При этом могут быть заданы ограничения и на значения атрибутов соответствующих отношений.

Например, запрос "Найти методы исследования, которые использовались для обработки деловых писем на русском языке в проектах в период с 1998 по 2005 год" будет выглядеть следующим образом:

Понятие "Метод исследования":
Отношение "Применяется к":
Понятие "Деловое письмо"
Атрибут "Язык" = "русский"
Отношение "Использует метод":
Понятие "Проект"
Атрибут "Дата начала": (>= 1998) & (<=2005)
Атрибут "Дата окончания": (>= 1998) & (<=2005)

Поисковые запросы задаются через специальный графический интерфейс, управляемый онтологией портала знаний. При выборе пользователем понятия автоматически генерируется поисковая форма, в которой можно задать ограничения на значения атрибутов объектов выбранного понятия, а также на значения атрибутов объектов, связанных с данным объектом ассоциативными отношениями.

Заключение

В докладе представлен подход к организации содержательного доступа к информационным ресурсам по компьютерной лингвистике путем построения специализированного (тематического) Интернет-портала.

Портал представляет знания об основных разделах компьютерной лингвистики, о ее предмете и объектах исследования, используемых в ней моделях и методах, разработанных системах, алгоритмах и лингвистических ресурсах, а также информацию об ученых, сообществах, организациях, включенных в процесс исследования по компьютерной лингвистике и о выполняемых проектах в этой области. Таким образом, пользователи портала имеют доступ не только к информационным текстовым ресурсам по компьютерной лингвистике, но и к ресурсам, представляющим реальные прикладные системы, технологии и программные продукты для обработки ЕЯ, лингвистические ресурсы и базы данных.

Для целостного представления знаний и данных по компьютерной лингвистике их систематизация и структуризация выполнены на основе онтологии. Благодаря этому, вся информация на портале представлена в виде сети взаимосвязанных информационных объектов.

Доступ к знаниям и данным портала осуществляется путем навигации по дереву понятий онтологии и информационному пространству портала, а также через средства содержательного поиска.

Портал знаний по компьютерной лингвистике разработан и доступен по адресу http://speedy.iis.nsk.su/cl/. При его создании использовалась технология, разработанная в ходе построения портала знаний по археологии [5, 6]. Для портала было разработано представительное ядро онтологии компьютерной лингвистики, которое на данный момент включает около 130 базовых понятий. В настоящее время выполняется информационное наполнение портала.

Ближайшей целью авторов является доработка онтологии компьютерной лингвистики, сбор и интеграция в информационное пространство портала новых лингвистических ресурсов.

Список литературы:

1. Guariano N., Giaretta P. Ontologies and Knowledge Bases. Towards a Terminological Clarification // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, 1995. P.25–32.
2. Боровикова О.И., Загорулько Ю.А. Организация порталов знаний на основе онтологий // Компьютерная лингвистика и интеллектуальные технологии: Труды международного семинара “Диалог 2002” (Протвино, 6-11 июня 2002 г.). М.: Наука, 2002. Т.2, С.76–82.
3. Borovikova O., Bulgakov S., Zagorulko Y., Sidorova E. Ontology-based approach to development of adjustable knowledge internet portal for support of research activity // Bulletin of NCC. Novosibirsk: NCC Publisher, 2005. Ser.: Computer Science. Is. 23, P.45–56.
4. Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции “Диалог 2005” (Звенигород, 1-5 июня 2005 г.). М.: Наука, 2005. С.65-70.
5. Загорулько Ю.А., Боровикова О.И. Технология построения онтологий для порталов знаний по гуманитарным наукам // Труды Всероссийской конференции с международным участием “Знания-Онтологии-Теории ”(ЗОНТ-07). Новосибирск, 2007. Т.1, С.191-200.
6. Андреева О.А., Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Циркин Б.Г., Холюшкин Ю.П. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии // Труды 10-й национальной конференции по искусственному интеллекту с международным участием КИИ'2006. М.: Физматлит, 2006. Т.3, С.832-840.