Опыт систематизации знаний и интернет-ресурсов для портала знаний по компьютерной лингвистике

Experience of systematizing knowledge and internet resources for a knowledge portal on computational linguistics

Соколова Е. Г. (minegot@rambler.ru)
Российский государственный гуманитарный университет, Москва

Загорулько Ю. А. (zagor@iis.nsk.su), **Кононенко И. С.** (irina_k@cn.ru)
Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск

В статье описывается опыт систематизации и интеграции знаний и интернет-ресурсов по компьютерной лингвистике в интернет-портал знаний. Рассматривается состав и структура объектов портала, место портала среди других каталогов по компьютерной лингвистике, опыт создания двуязычного словаря терминов по компьютерной лингвистике с использованием процедур автоматического извлечения терминов из текстов.

1. Введение

В ходе двухгодичного исследовательского проекта (2007–2008 годы), поддержанного РГНФ, создан интернет-портал знаний по компьютерной лингвистике (КЛ). Общая информация и проблемы, возникающие при описании в виде онтологии портала такой области как КЛ, рассмотрены нами в [1]. В данной статье мы обсуждаем различные аспекты проведенного исследования — принципы отбора информации, особенности портала и его место среди других каталогов по КЛ, классификацию и представление знаний и ресурсов, опыт создания двуязычного (англо-русского) словаря терминов по компьютерной лингвистике с использованием процедур автоматического извлечения терминов из текстов.

Принципы, которых мы придерживались, выбрав парадигму создания порталов знаний, заложенную работами [2, 3], состоят в следующем. 1. Главная цель проекта — представить российским исследователям всестороннюю картину КЛ, не претендуя на полноту в деталях. Тем самым мы стремимся отразить на портале достижения мировой КЛ, но при этом уделить особое внимание российской КЛ. 2. Российские исследования примерно на 15 лет позже, чем западные, начали переходить в стадию технологий и, как правило, не имели финансирования от организаций, требующих общедоступности результатов исследований, в связи с чем, в отличие от западных, редко завершались созданием ресурсов и функциональных систем. При этом отечественные исследования часто основывались на глубоких подходах и содержали интересные идеи, которые могут быть полезны сейчас или в будущем, поэтому нам хотелось включить и кратко охарактеризовать и такие отечественные публикации и материалы. 3. Из имеющихся в Интернете каталогов хотелось взять, прежде всего, «живые» системы и ресурсы, а также связанные с русским языком. 4. В начальной реализации портала язык описания его информационных объектов (проектов, методов и т. д.) — русский. Английский используется для указания исходных английских названий проектов, методов и моделей, например, «Расширенные сети переходов (Augmented Transition Networks (ATN))»; 5. Объекты онтологии сопровождаются только краткими описаниями, содержащими самую общую информацию об объекте. 6. Принципиальным требованием к таким проектам, как портал знаний, является наличие обратной связи с пользователями.

2. Границы описываемой порталом области исследований

В [1] мы показываем, что из нескольких терминов, соответствующих описываемой порталом области исследований, таких как прикладная

лингвистика, автоматическая обработка естественного языка, автоматическая обработка текстов, наиболее адекватен термин «компьютерная лингвистика», подчеркивающий тот факт, что компьютер является необходимой составляющей этой науки. Но недостаточной. Относится ли к КЛ оцифрованный и показанный на экране компьютера традиционный словарь или Грамматический словарь А. А. Зализняка, изданный в виде книги? Судя по тематике докладов на конференциях Диалога, на оба вопроса дается положительный ответ. В первом случае отнесение к КЛ оправдывается тем, что при формировании оцифрованного словаря используются компьютерные методы представления словарей и некоторые элементы формализации содержания словаря, хотя в целом оно остается не формализованным и ориентированным на человека. Во втором случае мы имеем дело с печатным изданием, но оно содержит формализованное описание русской морфологии, которое после введения в компьютер используется автоматизированными системами[1] морфологического анализа и синтеза русского языка. Если использование компьютера не является необходимым, то где граница между лингвистикой и КЛ? Граница есть, так как КЛ имеет свои собственные объект и предмет — преобразования текстов и звучащей речи[2], — которыми традиционная лингвистика не занималась и не занимается. Цели, которые ставятся в задачах преобразования текстов и речи, тем более выходят за рамки традиционной, преимущественно описательной, лингвистики, поэтому раньше КЛ называлась «прикладной». Современная КЛ имеет обширные пересечения с теоретической лингвистикой, искусственным интеллектом и математикой. КЛ использует теоретические достижения лингвистики для построения моделей языка в действии и преобразования текстов и речи, а сама дает методическую базу и ресурсы для проверки гипотез теоретической лингвистики. КЛ пересекается с ИИ в области обработки знаний и информации, выраженной на ЕЯ, использует методы математики для обработки текстов и речи в прикладных системах.

3. КЛ в зеркале онтологии портала знаний по КЛ

Понятия онтологии предметной области «Компьютерная лингвистика» организованы в 5 иерархий «общее-частное»: Объекты исследования, Предметы исследования, Разделы науки, Методы исследования, Научные результаты. Понятия в иерархиях связываются между собой посредством ассоциативных отношений.

Объекты исследования: Речевое произведение (РП) как объективная форма существования и использования естественного языка в виде Текста или Устной речи и Языковые единицы в составе РП, соответствующие различным языковым уровням: Синтаксические, Лексические, Морфологические и Фонетико-фонологические единицы. Для представления связи между целостными РП и их структурными единицами используется отношение «Включение».

Общеметодологический термин «объект исследования» ориентирован на традиционную науку и не совсем точен для КЛ, точнее было бы «объект моделирования». Особенность КЛ состоит в том, что собственно исследованием определенных лингвистических единиц она не занимается, но занимается созданием ресурсов лингвистических единиц — формализованных баз и корпусов, представляющих совокупности таких единиц. Организуя базы и корпуса, КЛ использует достижения традиционной лингвистики для систематизации и разметки единиц. Размеченные корпуса и компьютерные базы, с одной стороны, служат исходным материалом для настройки систем КЛ (например, при машинном обучении), а с другой стороны, позволяют верифицировать результаты лингвистических исследований. Объекты КЛ включают тексты и звучащие отрезки речи, которые обычно не являлись объектами систематического описания в лингвистике. В частности, структура текста как объект такого описания возникает в рамках КЛ для моделирования структуры текстов в системах генерации текстов.

Предмет исследования — аспект в исследуемом/моделируемом материале, на который направлена научная деятельность. Предметом исследования в КЛ являются 1. Процессы, связанные с функционированием языковых единиц в коммуникации. Среди них выделены процессы Анализа речи, Синтеза Речи, Анализа текста и Синтеза текста. В модели процессов выделены подпроцессы, которые относятся к уровням языка. Так, например, класс понятий Анализ текста представлен в иерархии подклассами: Сегментация текста, Морфологический анализ, Синтаксический анализ, Семантическая интерпретация, Анализ дискурса. Процессы анализа и синтеза имеют разный состав и не рассматриваются как обратимые. 2. Прикладные процессы, имеющие практическую ценность, отвечающие

---

[1] При подготовке Грамматического словаря русского языка А.А. Зализняк инициировал создание «Обратного словаря русского языка» с использованием методов машинной обработки материала, без которого работа над Грамматическим словарем не могла бы быть эффективно закончена.

[2] Ср. определение Автоматической Обработки Текстов, которое мы цитировали в [1]: «преобразование текста на искусственном или естественном языке с помощью ЭВМ» (с. 14, В.М. Андрющенко). Другими науками, порожденными компьютером, являются Искусственный Интеллект (Artifitial Intelligence), Вычислительная математика и Информатика (Computer science).

определенному социальному запросу, к которым относятся машинный перевод, автоматическое реферирование, идентификация говорящего по голосу и многое другое.

Методы так же, как и объекты, относятся скорее к способам моделирования, а не к способам исследования, хотя такие традиционные лингвистические методы, как, например, компонентный анализ, также включены в иерархию. Иерархия методов составляет центральную иерархию портала, так как КЛ является по сути методологической наукой. К методам отнесены Средства представления знаний, Грамматические формализмы, Методы теоретической лингвистики, Формальные механизмы и методы обработки ЕЯ, Методы оценки работы алгоритмов и систем. Теории языка в КЛ также носят методологический характер, т. е. объясняют не устройство систем конкретных единиц языка, а способы моделирования языковых средств для использования автоматическими системами. Такими теориями — моделями, породившими компьютерные системы, — являются структурная модель «Смысл-Текст» И. А. Мельчука, Ю. Д. Апресяна (компьютерные системы ЭТАП-2, RealPro и др.) и функциональная модель М. А. К. Хэллидея (компьютерные системы Penman, KPML, AGILE и др.). Модели Н.Хомского также представлены в иерархии методов как структурные модели, однако, если стандартную синтаксическую теорию Н. Хомского естественно рассматривать в рамках КЛ, где она повлияла на создание других уровневых моделей, в частности, российских (И. А. Мельчука, Ю. С. Мартемьянова), то теория GB перешла в ранг лингвистических теорий, позволяющих объяснять синтаксические явления.

В основе иерархии Разделов КЛ лежит классификация базовых теоретических и прикладных направлений КЛ: Моделирование языка и языковой деятельности (с разделами Автоматическая обработка текста (АОТ), Речевые технологии (РТ), Формализация описаний языковых средств и свойств речевых произведений) и Создание прикладных систем. В зависимости от направления моделирования (анализ или синтез) в первых двух разделах Моделирования языка и языковой деятельности выделены, соответственно, подклассы Понимание текста и Генерация текста, Распознавание речи и Синтез речи. В зависимости от объекта обработки (текст или звучащая речь), Прикладные системы разделены на Прикладные системы АОТ и Прикладные системы РТ.

Научные результаты представлены следующими классами: Технологии и программные продукты, Прикладные системы, Лингвистические ресурсы. Последний класс делится на такие классы, как Словари, Корпуса и Лингвистические БД. Класс Лингвистические БД, в свою очередь, разделен на Грамматические, Лексико-семантические, Семантико-синтаксические и Синтаксические ресурсы, а также Морфологические БД и Речевые БД. Корпуса разделяются на Корпуса текстов и Речевые корпуса.

4. Место портала по КЛ среди других интернет-каталогов

В Интернете имеются каталоги разработок и публикаций по КЛ, кратко рассмотренные нами в [4]. Наиболее крупный зарубежный каталог LINGUIST List [5] послужил прототипом для сайта «Российская лингвистика (RUSLING)» [6], созданного в Отделении лингвистических исследований ВИНИТИ РАН около 20 лет назад. Сайт «Лингвистика в России: ресурсы для исследователей», создан в феврале 2006 года по инициативе НИВЦ МГУ им. М. В. Ломоносова и ГОУВПО «Казанский государственный университет им. В. И. Ульянова-Ленина» [7]. Особенность Портала по КЛ по сравнению с этими каталогами состоит в том, что на нем представлена более узкая предметная область — КЛ, а информация структурирована в соответствии с онтологией предметной области. Полезными для КЛ являются такие источники, как Кругосвет, где имеются статьи, отражающие современное состояние лингвистики, например, статья, описывающая понятие дискурса, и интернет-энциклопедия Википедия, в которой можно найти полезную информацию о моделях, методах, интернет-ресурсах, персонах и организациях современной КЛ.

В Интернете представлена и более узко специализированная информация по отдельным направлениям КЛ. В качестве примера можно привести российский сайт «Речевые технологии» [8], всесторонне охватывающий теоретические и прикладные аспекты развития данного направления (технологии, программные средства, коллективы разработчиков, конкретные системы и т. п.).

Наиболее полными, точными и долговечными являются узкоспециализированные каталоги, поддерживающиеся западными исследователями, например, каталог систем генерации текстов [9]. В нем содержится информация обо всех западных системах и проектах по созданию таких систем, всего на момент написания данной статьи 383 системы. Например, для проекта AGILE приводится следующая информация:

Имя системы: AGILE
Разработчики: Krujiff, Korbayová, Teich, Hartley, Bateman, Sharoff, Scott, Staykova, Sokolova
Даты разработки: 1999–2001
Языки: Bulgarian, Czech, Russian
URL (if available) http://www.itri.bton.ac.uk/projects/agile who's who
Построен на основе: KPML
Описание: AGILE is a tool which allows a technical author to specify, in a non-linguistic representation, the 'content' of different tasks that can be performed by users of CAD-CAM software. The AGILE system can then automatically express these content specifications in styles appropriate to different sections of a CAD-CAM manual (procedures, ready reference ...) in Bulgarian, Czech and Russian. The generated texts are displayed in a browser as hyperlinked documents. No expertise in knowledge representation is required, although some training with the interface is needed. The system has been evaluated and the results are described in the relevant project deliverables.
Ссылки на публикации по проекту (3 ссылки).

Задача создателей этого каталога облегчается тем, что описываемые системы создавались и создаются преимущественно в рамках целенаправленно финансируемых научных проектов, имеющих конкретный срок разработки от 2 до 5 лет и направленных на создание одной конкретной системы или среды.

В отличие от узкоспециализированных каталогов, портал по КЛ охватывает все типы метапонятий, определенных прототипом портала, в соответствии с чем на портале представлена как информация собственно по КЛ (теории, методы и т.д.), так и информация по автоматическим системам КЛ, которая разнесена по онтологическим классам, а соответствующие им объекты связаны в единую сеть содержательными ассоциативными отношениями, обеспечивающими переходы по сети от одного элемента информации к другому.

Так, информация о проекте AGILE распределена по онтологическим классам следующим образом: Деятельность → Проект → проект AGILE; Объект исследования → Речевое произведение → Текст → Инструкция; Раздел науки → Моделирование языка и языковой деятельности → Автоматическая обработка текста → Генерация текста; Научные результаты и продукты → Прикладная система → система AGILE; Персона → <список исполнителей>; Организация → <список организаций-участников>.

Объекты, соответствующие этим классам, связаны отношениями: «Исследует объект», «Использует результат», «Результат деятельности», «Персона-Участник деятельности», «Публикация о деятельности» и т. д. Например, отношение «Использует результат» связывает проект AGILE со средой для разработки многоязыковых генераторов KPML, который, в свою очередь, позволяет выйти на персону-автора данного ресурса и т. д.

Портал знаний не только обеспечивает гибкое представление информации по КЛ, но и предоставляет пользователям удобные средства поиска и навигации по ней.

Для навигации по контенту портала используется дерево понятий онтологии. При выборе в этом дереве определенного узла пользователь получает список соответствующих ему информационных объектов. Если на объектах выбранного понятия задано отношение включения, по желанию пользователя этот список может быть представлен в виде дерева. Информация о конкретном объекте и его связях отображается в виде html-страницы, при этом объекты и интернет-ресурсы, связанные с данным объектом, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.

Портал предоставляет пользователю два вида поиска: простой поиск и расширенный поиск. Простой поиск позволяет находить объекты, в значениях атрибутов которых содержится строка поиска. Расширенный поиск предоставляет пользователю возможность задания запроса в терминах предметной области портала. При этом пользователь может указать не только объекты какого понятия и с какими свойствами он хочет найти, но и задать ограничения на значения атрибутов объектов, связанных с искомым объектом.

Таким образом, портал может служить основой для конечных пользователей при поиске проектов, интернет-ресурсов и методов исследований по КЛ. С другой стороны, на нем могут базироваться библиотечные классификации данной области, которые в настоящее время остаются неразработанными.

В связи с масштабностью поставленной задачи на портале представлены не все проекты, лингвистические группы и исследователи, не полностью охвачены модели и методы, используемые в КЛ. Поэтому работа по развитию онтологии КЛ и наполнению контента портала новыми данными и ресурсами будет продолжена.

5. Словарь терминов КЛ

Для поддержки автоматизации сбора интернет-ресурсов и новостей, относящихся к области КЛ, а также обеспечения визуализации контента портала и поиска в нем информации на разных языках было разработано несколько словарей. Основой этих словарей стали английский, русский и англо-русский словники, построенные в ходе выполнения проекта в 2007–2008 годах.

В качестве материала для создания этих словников были сформированы два корпуса текстов — английский, включающий учебник под ред. Р. Митков [10] и обзор под ред. Р. Коула [11], и русский, включающий труды конференций «Диалог» за последние три года (2006, 2007, 2008), взятые с сайта конференции Диалог, ввиду отсутствия фронтальных обзоров[3] (о чем мы писали в статье [1]). Оба корпуса были обработаны с помощью технологии автоматического создания терминологической базы по текстам предметной области, авторами которой являются Н. В. Лукашевич и Б. В. Добров [12], в результате чего были получены списки русских и английских терминов, в которые вошли слова и правильные словосочетания (в основном, пары слов —

[3] Учебники по КЛ и близким областям, изданные в России за последние два десятилетия: Шемакин Ю. И. Начала компьютерной лингвистики. М.: Изд-во МГОУ А/О «Росвузнаука», 1992. — 114 с.; Прикладное языкознание. Учебник. (ред. А.С.Гердт). СПб., 1996.; Баранов А. Н. Введение в прикладную лингвистику. Серия «Новый лингвистический учебник». М.: Эдиториал УРРС. 2001; Леонтьева Н.Н. Автоматическое понимание текстов. Системы. Модели. Ресурсы. М.: Academia, 2006, — не дают стройной картины направления, которую мы видим в западных учебниках и обзорах этого же периода.

согласованные Прил+Сущ и Сущ+Сущ в родительном падеже, а также порожденные на их основе трехкомпонентные словосочетания). Русский список был пополнен с помощью технологии [13], примененной к текстам, содержащим определения понятий, атрибутов, доменных значений и объектов, составленных экспертом при описании онтологии КЛ. В результате объем русской терминологической базы достиг 13 тыс. слов, английской — 15,5 тыс. слов. После просмотра и редактирования экспертом количество терминов существенно сократилось и составило, соответственно, 4801 русских и 4972 английских термина. На основе этих списков с учетом статистических данных был создан двуязычный (англо-русский и русско-английский) словарь по КЛ.

Сравнение полученных русских и английских списков показало, что многие единицы в них, с точки зрения эксперта, не являются терминами, кроме того, они содержали не вполне коррелированные наборы понятий. В результате в словарь вошла часть терминов, для которых было установлено межъязыковое соответствие путем сопоставления английского и русского терминов из списков, а другая часть получена путем перевода английских терминов на русский язык.

Полученный словарь включает около 1900 англоязычных и русскоязычных терминов и их переводов (английских — 1866, русских — 1875, общее количество связей — 2199: связей больше ввиду синонимии; все непереведенные слова были автоматически удалены из словаря). Этот результат свидетельствует об ограниченности проведенного эксперимента (нужны значительно более объемные корпуса текстов) и о неполной сопоставимости исследований по КЛ в России (как они представлены в Диалоге за последние три года) и за рубежом.

Кроме того, на основе английской и русской терминологии было разработано два предметных словаря, содержащих морфологическую, статистическую и семантическую информацию. Эти словари использовались для настройки модулей портала, отвечающих за автоматизацию наполнения его контента и сбора новостных сообщений по тематике КЛ.

6. Заключение

В настоящее время портал знаний доступен по адресу http://uniserv.iis.nsk.su/cl. Его контент включает более 600 интернет-ресуров, около 2000 информационных объектов, связанных примерно 4000 отношениями. Пользователь может видеть не только иерархию «общее-частное», заданную на понятиях онтологии, но и иерархии «часть-целое», заданные на информационных объектах.

Работа по развитию онтологии КЛ и наполнению контента портала новыми данными и ресурсами будет продолжена. Разработанный англо-русский словарь будет дополнен и станет базой для визуализации контента портала и поиска в нем информации на двух языках — русском и английском.

Обратная связь с пользователем на данный момент осуществляется через адрес электронной почты, указанный на сайте в разделе «О портале». Планируется организация специального форума для обсуждения онтологии и контента портала.

Авторы благодарят О. Ф. Кривнову за помощь в систематизации и работе с терминологией для разделов по обработке речи, а также Н. В. Лукашевич и Б. В. Доброва за проведение эксперимента по извлечению терминологии.

Литература

1. *Соколова Е. Г., Кононенко И. С., Загорулько Ю. А.* Проблемы описания компьютерной лингвистики в виде онтологии для портала знаний // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008» (Бекасово, 4-8 июня 2008 г.). М.: РГГУ, 2008. Вып. 7 (14), С. 482–487.
2. *Боровикова О. И., Загорулько Ю. А., Загорулько Г. Б., Кононенко И. С., Соколова Е. Г.* Разработка портала знаний по компьютерной лингвистике // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием КИИ-2008 (г. Дубна, Россия). М.: ЛЕНАНД, 2008. Т. 3, С. 380–388.
3. *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.
4. *Загорулько Ю. А., Боровикова О. И., Загорулько Г. Б.* Портал знаний по компьютерной лингвистике: содержательный доступ к лингвистическим информационным ресурсам // Компьютерная лингвистика и интеллектуальные технологии. Электронные публикации Международной конференции «Диалог-2008» (http://www.dialog-21.ru/dialog2008/materials/html/Zagorulko.htm)
5. *LINGUIST List* (http://linguistlist.org/)
6. «*Российская* лингвистика (RUSLING)» (http://rusling.narod.ru)
7. «*Лингвистика* в России: ресурсы для исследователей» (http://uisrussia.msu.ru/linguist/index.jsp)
8. *Портал* «Речевые технологии» (http://speechsoft.ru/)
9. «*John Bateman and Michael Zock's* List of NLG systems» http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.htm
10. *Mitkov Ruslan* (ed.) The Oxford handbook of computational linguistics // N.Y.: Oxford university press, 2003.
11. *Cole Ronald* (ed.) Survey of the state of the Art in Human Language Technology // 1996. (http://cslu.cse.ogi.edu/HLTsurvey/).
12. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции», 2003. С. 201–210.
13. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008» (Бекасово, 4–8 июня 2008 г.). М.: РГГУ, 2008. Вып. 7 (14), С. 475–481.