Разработка русского WordNet

Сухоногов А.М.

Петербургский университет
путей сообщения
ASukhonogov@rambler.ru

Яблонский С.А.

Петербургский университет
путей сообщения, ЗАО
"Руссикон"
serge_yablonsky@hotmail.com

Аннотация

В статье описывается реализация лексико-семантической базы данных типа WordNet для русского языка, методика ее построения, вопросы интеграции с другими реализациями WordNet и представление WordNet в формате OWL (Ontology Web Language).

1 Реализации WordNet

1.1 Развитие WordNet словарей

Работа над словарем WordNet [1] английского языка начата в Принстонском университете (США) в начале 80-х годов и продолжается до настоящего момента. Сейчас доступна версия 2.0. этого словаря. Существующая версия WordNet охватывает общеупотребительную лексику современного английского языка – более 120 тысяч слов. Широкое распространение этот словарь получил благодаря его свободной доступности для научных и исследовательских целей.

В период с марта 1996 по сентябрь 1999 года при финансировании Европейской комиссии был создан многоязычный вариант WordNet – EuroWordNet [2]. Эта лексическая система объединила в себе WordNet словари английского, датского, испанского, итальянского, немецкого, французского, чешского и эстонского языков, за основу был взят Принстонский WordNet версии 1.5. В 2004 году завершается работа над проектом BalkaNet, объединяющем греческий, болгарский, турецкий, чешский, французский, румынский и сербский языки. Все национальные версии WordNet связаны с исходным WordNet и между собой через специальный ILI–индекс. Традиционный подход предполагает использование при построении WordNet словаря специализированных систем разработки, например VisDic (проект BalkaNet) [8]. Словари EuroWordNet являются коммерческим продуктом.

В настоящее время словари WordNet могут применяться в системах информационного поиска (information retrieval), вопросно-ответных системах (Q&A systems), в системах машинного перевода (machine translation) и при решении задачи определения значения слов (WSD - word-sense-disambiguation).

1.2 Проекты словарей WordNet русского языка

В настоящее время известно о нескольких реализациях WordNet подобных лексических баз данных для русского языка:

1. Проект RussNet, разрабатывается с 1999 года на филологическом факультете СПбГУ [4].
2. Проект тезауруса RuThes, используемого в УИС РОССИЯ [5]. Закрытый коммерческий ресурс.
3. Проект русского WordNet компании «Новософт»[6]. Закрытый коммерческий ресурс.

Методика и принципы построения словаря проекта RussNet [7] ориентированы на длительный процесс разработки ресурса группой лингвистов без какой-либо автоматизации процесса построения и связи с исходным WordNet. Два других проекта невозможно оценить из-за их закрытости, хотя в последнем используется небольшой англо-русский словарь Миллера для автоматизации построения ресурса.

Рассматриваемая в данной работе реализация русской версии WordNet позволяет получить ядро словаря в меньший срок за счет использования доступных словарей и автоматизации процедур построения и редактирования словаря. Ставится задача получения русской версии WordNet сопоставимой по числу лексических единиц с английской версией.

Для этих целей разработана методика, включающая набор алгоритмов и процедур их проверки. Разработанные методы, позволяют значительно сократить время разработки за счет более эффективного использования существующих ресурсов и автоматизации процесса построения словаря WordNet на их основе. Большое внимание уделено вопросу интеграции с другими лексическими ресурсами. Однако для повышения качества получаемого таким образом словаря его ручная доработка на каждом этапе построения неизбежна.

Труды 6ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2004, Пущино, Россия, 2004.

2 Методика разработки wordnet-словаря

2.1 Электронные ресурсы

Для построения русского WordNet используются лингвистические ресурсы компании «Руссикон» [9,10] и словари, свободно распространяемые в Internet, например, [11]. Научный коллектив из сотрудников ПГУПС (каф. ИВС) и компании Руссикон под руководством Яблонского С.А. выиграл в 2003 г. конкурс издательства Oxford Press на лучший исследовательский проект по использованию словарей Oxford Press. В настоящее время издательства Oxford Press предоставило для создания русской версии WordNet XML версии следующих словарей:

1. Oxford Russian Dictionary.
2. New Oxford Dictionary of English, 2nd Edition.
3. New Oxford Thesaurus of English.

Эти ресурсы используются при автоматизированном построении межъязыкового индекса (ILI–Inter-lingual-index) русско-английского WordNet.

2.2 Этапы построения

Основной целью нашего проекта является построение русско-английского WordNet, включающего лексику русского и английского языков. Разработка такого варианта словаря включает два этапа – построение русского WordNet, описывающего лексику русского языка и объединение полученного WordNet с последней реализацией Princeton WordNet с помощью ILI.

Для просмотра и редактирования словаря разработан редактор – TenDrow, позволяющий просматривать и редактировать словарные статьи WordNet и иерархии их отношений (строятся деревья гипонимии (родовидовые отношения) и меронимии (отношения часть-целое)). Редактор используется для «чистовой» обработки словаря. Как показала практика разработки и построения словаря, наиболее эффективным является редактирование специально подготовленных текстовых файлов и набор утилит для внесения изменений в базу данных. Для каждого этапа построения словаря формируется набор таких файлов и средств их обработки.

Базовой структурной единицей, формирующей словарную статью WordNet, является синонимичный ряд – «синсет», объединяющий лексемы со схожим значением. Каждый синсет представляет некоторое значение, понятие языка. Для каждого синсета определяется толкование, уточняющее это значение и примеры употребления лексем синсета в некотором контексте.

На первом этапе анализируется толковый словарь [9, 10], из него выделяются значения слов с толкованиями – прототипы синсетов. Прототипы синсетов включают от 1 до 6 слов-синонимов, полученных по пометам, определяющим эквивалентность значений слов с словарных статьях толкового словаря. Для всех полученных таким образом лексем определяется полная парадигма – производится привязка статей к грамматическому словарю, при этом лексемы различаются не только по частям речи, но и по другим признакам, например, одушевленности. Полученный словарь пересекается со словарем синонимов - тезаурусом «Руссикон», синсеты дополняются синонимами и формируется дерево гипонимии, соответствующее структуре тезауруса.

На втором этапе полученный словарь пересекается с Принстонским словарем WordNet. Для реализации такого пересечения используются Оксфордские словари.

Мы пытаемся последовательно воспроизвести отношения синсетов, которые определяются деревьями гипонимии и меронимии исходного WordNet. Осуществляется обход этих деревьев «в ширину». Для каждого синсета WordNet предпринимается попытка найти синсет среди множества синсетов-прототипов русского словаря, полученных на первом этапе. Такой поиск предполагает использование не только словника синсетов и частотного словаря, как в других реализациях [12] – этого оказывается недостаточным, поскольку много синсетов у корней деревьев состоят всего из одной леммы, имеющей множество лексем, например, *'make'*. Для синсетов словарей анализируются толкования и примеры употребления – они нормализуются, переводятся и сравниваются. Эта косвенная информация позволяет значительно повысить качество автоматически получаемого результата. Последовательный обход деревьев отношений позволяет воспроизвести их структуру в русском WordNet (где это возможно) и определить отношения тождества (EQ-отношения), составляющие ILI [2].

2.3 Структура данных

Структура данных разработана с учетом существующих реализаций WordNet и форматов представления этого словаря. Значительное влияние на структуру оказала и методика построения русского WordNet. В нашем проекте для хранения данных используется РСУБД Oracle 9i. Существует множество реализаций структуры WordNet для РСУБД, например[13], наша реализация имеет ряд отличий, рис.1.

В состав словаря включен грамматический словарь (GramTree, GramProp, WordForm) - для каждой леммы русского WordNet определена полная парадигма. Также определяются словообразовательные отношения между леммами (WordBuild). В Принстонском WordNet это реализовано процедурно, флективный русский язык сложнее и имеет массу исключений.

Толкование и примеры употребления лексем (Saying_Idiom) в нашей реализации привязываются не только к синсетам, но и к отдельным лексемам,

что связанно с тем, что в основе WordNet лежит толковый словарь.

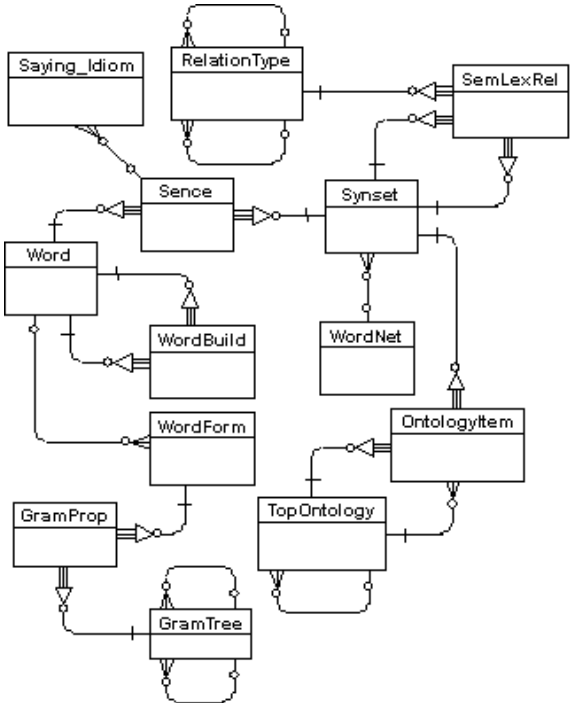

Рис.1. ER-диаграмма реализации WordNet

Существенное различие – в реализации отношений между лексемами при реализации таких отношений, как антонимия. Наша реализация позволяет определять такие отношения как между синсетами, так и между лексемами. Это необходимо для поддержания совместимости с XML-форматом редактора VisDic, где отношения между лексемами не определяются. Такие отношения реализуются за счет последовательной нумерации всех лексем синсета и включения этих номеров в состав отношения SemLexRel.

Синсеты могут привязываться к TopOntology – онтологии, выделенной из проекта EuroWordNet[2] (63 класса). Структура позволяет загружать вместо/вместе TopOntology и другие классификаторы, например рубрикатор ГАСНТИ и/или иерархию WordNet Domains [14].

В состав отношения RelationType добавляются определения EQ-отношений тождества, необходимые для построения ILI, а каждый синсет привязывается к своему словарю WordNet (WordNet).

Та же структура в нотации UML описана в [15].

2 WordNet в технологиях SemanticWeb

Для представления русского WordNet и его интеграции с другими программными системами разработана OWL/RDFS-схема WordNet

В настоящее время реализованы процедуры экспорта-импорта данных лексической базы данных в формат Princeton WordNet (версии 2.0), XML-формат редактора VisDic (версии 1.3.36) и собственный OWL-формат.

Таким образом, WordNet может использоваться как один из компонентов технологии W3C – SemanticWeb, опирающейся на открытые стандарты. Предложенная схема соответствует основным рекомендациям W3C-консорциума [16,17,18]. OWL-схема проверена на корректность с помощью RDF-анализатора Jena [19].

3 Текущее состояние

В настоящее время словарь находиться на стадии проверки, редактирования и определения ILI-индекса.

Статистика по леммам и синсетам текущей реализации русского WordNet приводиться в таблице.1.

Таблица 1

	Лемм	Синсетов
Существительные	46112	59880
Глаголы	30185	44078
Прилагательные	21936	28833
Наречия	6296	5868

Для просмотра и редактирования словаря создан редактор TenDrow. Для просмотра данных словаря - Internet/Intranet реализация. Данные будут размещены для публичного просмотра после завершения проверки:
URL: http://www.pgups.ru/WebWN/wordnet.uix

Реализация OWL/RDFS-схемы WordNet доступна:
URL: http://www.russicon.ru/wordnet/wn-russicon.htm.

4 Заключение

Рассмотренная система предназначается для создания и редактирования широкого класса тезаурусов и близких к ним структур. Реализация набора интерфейсов к этим системам позволяет использовать их как самостоятельные приложения – лексикографическая система WordNet и система классификаторов, так и включать их в состав более сложных систем.

Литература

[1] C.Fellbaum (ed.), WordNet: An Electronic Lexical Database, MIT Press, 1998.

[2] P.Vossen. Building a multilingual database with wordnets for several European languages. http://www.illc.uva.nl/EuroWordNet/

Труды 6ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2004, Пущино, Россия, 2004.

[3] C.Fellbaum, P.Vossen. The Global WordNet Association. http://www.globalwordnet.org/
[4] Сайт проекта RussNet http://www.phil.pu.ru/depts/12/RN/index_ru.shtml
[5] Портал УИС «Россия» http://www.cir.ru/
[6] Компания «Новософт», Новосибирск. http://www.novosoft.ru
[7] И.В. Азарова, А.А. Синопальникова, М.В. Яворская Принципы построения wordnet-тезауруса RussNet. Труды конференции Диалог-2004, http://www.dialog-21.ru/Archive/2004/Sinopalnikova.htm
[8] A.Horák, P.Smrž "VisDic – Wordnet Browsing and Editing Tool". P.Sojka, K.Pala, P.Smrž, C.Fellbaum, P.Vossen (Eds.): GWC 2004, Proceedings, pp. 136–141. Masaryk University, Brno, 2003
[9] S.A. Yablonsky,. Russicon Slavonic Language Resources and Software. In: A.Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) Proceedings First International Conference on Language Resources &Evaluation, Granada, Spain, 1998.
[10] A.M. Sukhonogov "Language resources usage for English-Russian WordNet development". Applied Natural language Processing – possible application for Semantic Web. Eurolan 2003. Student Workshop Proceeding. Hamburg University, Juli 2003.
[11] Словари, энциклопедии, справочники http://www.slovarik.ru/, http://www.artint.ru/projects/frqlist.asp.
[12] И.Г.Гельфенбейн, А.В. Гончарук и др. Автоматический перевод сети WordNet на русский язык. Материалы конференции Диалог-2003
[13] R.Bergmair WordNet ERD, 2002 http://wordnet2sql.infocity.cjb.net/model-overview.html
[14] B.Magnini, G.Cavaglia. «Integrating Subject Field Codes into WordNet». In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, pp. 1413-1418 Athens, Greece, 31 May - 2 June, 2000
[15] V.Balkova, A.Sukhonogov, S.Yablonsky "Russian WordNet. From UML-notation to Internet/Intranet Database Implementation". GWC 2004, Proceedings, pp. 31–38. Masaryk University, Brno, 2003. – 374 pp.
[16] RDF Vocabulary Description Language 1.0. RDF Schema. http://www.w3.org/TR/2003/PR-rdf-schema-200331215
[17] OWL Web Ontology Language. Guide. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-guide-20040210
[18] OWL Web Ontology Language. Reference. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-owl-ref-20040210
[19] HP Labs Semantic Web Research http://www.hpl.hp.com/semweb/

Труды 6ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2004, Пущино, Россия, 2004.

Russian WordNet development

A.Sukhonogov, S.Yablonsky

This paper describes the development of the first public web version of Russian WordNet and future parallel English-Russian and multiligual web versions of WordNet. It reviews the usage of Russian and English-Russian lexical language resources and software to process WordNet for Russian language and design of a database management systems for efficient storage and retrieval of various kinds of lexical information needed to process WordNet. The pilot Internet/Intranet version of described system is published at: http://www.pgups.ru/WebWN/wordnet.uix.