Интеграция тезаурусов RussNet и YARN

И.В. Азарова [1], П.И. Браславский [2], В.П. Захаров [1],
Ю.А. Киселев [2], Д.А. Усталов [3], М.В. Хохлова [1]

[1] Санкт-Петербургский государственный университет
i.azarova@spbu.ru, v.zakharov@spbu.ru, m.khokhlova@spbu.ru
[2] Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина
pbras@yandex.ru, ykiselev.loky@gmail.com
[3] Институт математики и механики им Н.Н. Красовского
Уральского отделения Российской академии наук
dmitry@eveel.ru

Аннотация

На сегодняшний день отсутствует большой открытый тезаурус русского языка под свободной лицензией. Предлагается проект интеграции двух электронных тезаурусов русского языка. Специфика конкретных ресурсов и особенности русского языка определяют оригинальность и научную новизну методов, используемых для объединения. Результатом проекта будет полноценный русскоязычный тезаурус, интегрирующий данные RussNet (40 тыс. слов и словосочетаний, 30 тыс. синсетов, 45 тыс. семантических связей) и YARN (120 тыс. словарных единиц, 46 тыс. синсетов, 30 тыс. иерархических отношений) с дополненными и отредактированными данными. Важным аспектом проекта является сочетание подходов на основе краудсорсинга и работы экспертов.

Ключевые слова: лексическая статистика, тезаурус, лингвистическая онтология, WordNet, интеграция словарных данных.

1. Введение

Большой общедоступный семантический словарь в электронном виде сегодня входит в стандартный набор необходимых инструментов и ресурсов для автоматической обработки текстов на конкретном языке (наряду с морфологическим анализатором, синтаксическим парсером, большим аннотированным корпусом и т. п.). Стандартный подход к организации таких ресурсов реализован в проекте Princeton WordNet (PWN), работа над которым

началась в 1986 г. Принципы разработки PWN, методы автоматического пополнения тезауруса, а также некоторые приложения на основе данного ресурса описаны в книге [1]. На сегодняшний день WordNet-подобные ресурсы созданы для многих языков. Их обзор можно найти в [2]. Однако до сих пор отсутствует большой открытый WordNet-подобный тезаурус русского языка высокого качества, распространяемый под открытой лицензией.

В настоящее время WordNet-подобными словарями называют лексические базы, построенные по базовым принципам проекта PWN. В частности, они состоят из синсетов (synset, от synonym set) — «смыслов», которые выражаются набором квазисинонимов. В свою очередь синсеты связаны между собой различными семантическими отношениями — гипероним-гипоним, меронимхолоним и др. В PWN входят значения существительных, глаголов, прилагательных и наречий. Текущая версия PWN содержит более 117 тыс. синсетов, которым соответствуют примерно 150 тыс. различных словарных входов (отдельных слов и фраз). PWN успешно используется для решения широкого круга задач: снятия лексической неоднозначности, автоматического реферирования, семантического поиска, классификации и кластеризации документов, обработки поисковых запросов, машинного перевода и т. д.

2. WordNet для русского языка

Компьютерный WordNet-подобный тезаурус русского языка RussNet появился в 1999 году в качестве инициативного проекта кафедры математической лингвистики СПбГУ [3]. RussNet состоит из иерархических структур значений четырех основных частей речи: существительных, глаголов, прилагательных и наречий. Единицами описания являются синсеты, в состав которых входят лексемы и устойчивые словосочетания, выражающие лексические значения, оформленные в русском языке. Между синсетами устанавливаются семантические отношения (гипонимические, меронимические, каузативные, пресуппозиционные), между лексемами — семантико-деривационные. В настоящее время в разной степени готовности создано около 30 тыс. синсетов и 45 тыс. семантических связей, заданных на 40 тыс. слов и словосочетаний. RussNet создается как оригинальная система лексикализованных понятий русского языка, а не как перевод тезауруса английского языка WordNet. Основными недостатками проекта RussNet являются (1) отсутствие онлайнового представления тезаурусных данных, (2) фрагментарное покрытие семантических областей; (3) отсутствие привязки синсетов RussNet к PWN, который является своеобразным словарем-посредником для представления содержания текстов на разных языках.

Принципиально другой стратегии построения электронного тезауруса, которая не продемонстрировала своей эффективности, придерживались разработчики Russian WordNet (RWN) из Петербургского государственного университета путей сообщения и ЗАО «Руссикон» [4] и из Новосибирского университета [5] (http://www.dialog-21.ru/Archive/2003/Goncharuk.htm). Эта стратегия заключается в полностью автоматическом переводе тезауруса английского языка PWN на русский.

Предлагаемая ЗАО «Руссикон» версия WordNet является принципиально параллельной англо-русской, т.е. синсеты PWN и отношения между ними переносятся на лексико-семантические варианты русских слов. Для построения и редактирования RWN предлагалось использовать существующие двуязычные словари и автоматические методы. К сожалению, результатов проекта нет в открытом доступе, что делает невозможным самостоятельный анализ его данных.

Результат работы представителей НГУ в формате PWN доступен в сети Интернет (см. http://www.wordnet.ru) — тезаурус содержит примерно 18 тыс. существительных, 6 тыс. прилагательных, 5,5 тыс. глаголов, 1,8 тыс. наречий. Существенным недостатком этого проекта является то, что часть его данных осталась на языке оригинала, то есть на английском, например, определения понятий, представленных в ресурсе.

Методология перевода Princeton WordNet на русский язык обеспечивает относительную быстроту заполнения лексикографических баз, более простую интеграцию в аналогичные многоязычные системы и подключение различных (в том числе толковых) электронных словарей. Однако существующие межъязыковые различия, т.е. асимметричные явления в лексике, препятствуют наложению семантической структуры одного языка на лексику другого, что существенно затрудняет не только автоматический перевод тезауруса PWN, но и ручной.

Тезаурус YARN (http://russianword.net/) разрабатывается в Уральском федеральном университете с 2013 года для задач автоматической обработки текста и информационного поиска. YARN включает в себя около 120 тыс. словарных единиц, 46 тыс. синсетов и 30 тыс. иерархических отношений [6]. Лексикографический подход YARN предполагает сочетание краудсорсинга с автоматическими методами построения тезаурусов. В рамках проекта разработаны онлайн-инструменты для коллективного редактирования тезауруса, а также автоматические методы подготовки «сырья» для построения ресурса на основе данных словарей и анализа корпусов текстов. Из-за ограниченности ресурсов в рамках проекта YARN наиболее полно представлены синсеты существительных; родовидовые отношения требуют дополнительной проверки.

Таким образом, несмотря на достаточно большое количество разработок электронных тезаурусов русского языка, на данном этапе приходится констатировать, что ни один из представленных проектов не доведен до конца и не может рассматриваться как полноценный русскоязычный WordNet-подобный тезаурус. В то же время наличие большого открытого электронного тезауруса русского языка для задач автоматической обработки текстов существенно способствовало бы развитию отечественных исследований и разработок в областях искусственного интеллекта, информационного поиска, компьютерной лингвистики и повысило бы эффективность и качество работы различных систем компьютерной обработки русского языка.

3. Новый подход: интеграция имеющихся ресурсов

Авторы статьи исходят из того, что русскоязычный WordNet-подобный тезаурус должен создаваться с учетом особенностей русского языка. Предлагается проект объединения двух ресурсов — RussNet и YARN. Главный результат реализуемого проекта — большой открытый тезаурус русского языка, построенный на основе двух указанных ресурсов, отредактированный и дополненный новыми данными. В результате интеграции итоговый ресурс будет обладать преимуществами каждого из оригинальных ресурсов: проверенное качество данных (полученных из RussNet) и готовые схемы представления данных, а также программный и графический интерфейс для работы с новым тезаурусом (из YARN). Тезаурус, полученный в результате объединения двух ресурсов, будет обладать более высокой полнотой лексических единиц и синсетов, а также семантических связей, по сравнению с нынешним наполнением каждой из его компонент. Дополнительная идея — эксперимент по комбинированию лингвистических принципов создания WordNet-подобных словарей и коллективного подхода к наполнению и редактированию лингвистических ресурсов. Данный тезаурус будет распространяться под лицензией CC BY-SA, подразумевающей максимально широкое использование создаваемого ресурса.

Однако для достижения результата необходимо выполнить интеграцию гетерогенных лексикографических данных: RussNet построен путем лексико-статистического лингвистического подхода, YARN создан путем краудсорсинга с дополнительным применением автоматических методов построения тезаурусов. Интеграция включает в себя согласование концептуальных оснований двух ресурсов, схем данных, разработку автоматических методов выравнивания и сравнения единиц тезаурусов; методики, сценариев и инструментов редактирования и пополнения объединенного ресурса.

Благодаря планируемому объединению, последующая работа над ресурсом, в том числе и в рамках описываемого проекта, связанная с наполнением его новыми данными, будет возможна как за счет работы экспертов, так и с использованием краудсорсинга.

Задача интеграции семантических ресурсов решается в рамках направления создания открытых связанных данных. Специфика конкретных ресурсов и особенности конкретного языка определяют оригинальность и научную новизну используемых методов.

4. Схема интеграции

В настоящее время проходит непростой этап согласования форматов передачи информации. Одновременно решается задача редактирования и пополнения данных, представленных в RussNet, поскольку в методике построения учитывался *фактор статистической реализации* значений слов заданной семантической области *в корпусе* современных текстов. В структуре синсетов YARN, напротив, достаточно часто представлены слова, имеющие «потенциальный» характер, т.е. они либо вовсе не встречаются в текстах НКРЯ и других корпусов, либо имеют единичное вхождение, причем не в том значении, которое приведено в синсете. В основном это касается

терминологических употреблений слов, архаизмов, диалектизмов или жаргонизмов. Следует заметить, что в методике построения PWN отмечалось, что словарь является нетерминологическим и содержит активно употребляющиеся слова современного языка [1], однако фактическое содержание синсетов PWN часто противоречит этому принципу.

Разрешение этого противоречия, вероятнее всего, пойдет по линии наиболее полного наполнения синсетов, в которых будут выделены нейтральное *слово-«доминанта»*, реализующее значение синсета частотно и стилистически независимо, ряд среднечастотных слов, имеющих ограниченную сочетаемость или стилистическую привязку, и практически неограниченный набор переносных употреблений слов, стилистически маркированных архаизмов, жаргонизмов и проч.

Таким образом, начальный этап интеграции включает согласование схем данных двух ресурсов, разработку интерфейса для автоматического преобразования данных, подготовку 3–4 пилотных областей значений слов разных частей речи для интеграции; разработку методики редактирования данных; представление данных RussNet в консолидированном формате; модификацию существующего инструмента редактирования YARN с учетом задачи интеграции ресурсов; тестирование системы на пилотных областях; первичное пополнение семантических областей; корректировку методики редактирования данных.

На рис. 1 представлена схема понятий Linked Data, используемых для представления элементов проектируемого тезауруса: понятий (skos:Concept), лексических значений (lemon:LexicalSense), лексем (lemon:LexicalEntry), определений (lemon:SenseDefinition), примеров употребления (lemon:UsageExample) и словоформ (lemon:Form) [7].

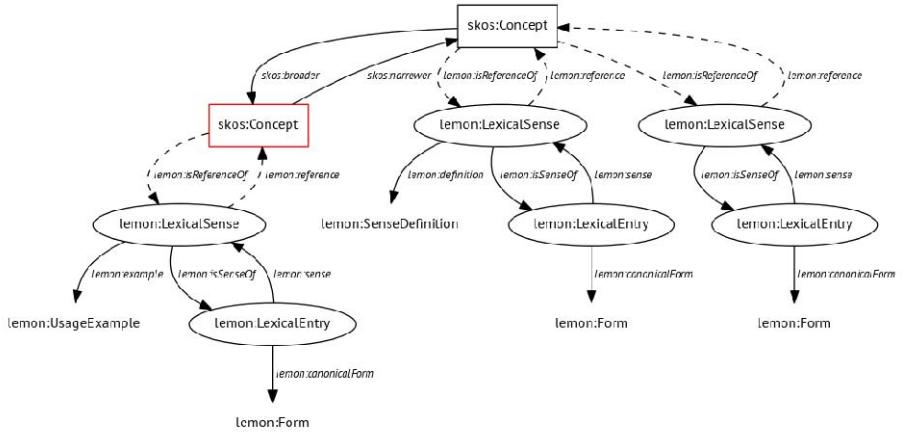

Рис. 1. Схема для представления элементов предлагаемого тезауруса на основе понятий Linked Data

5. Заключение

Создаваемый тезаурус может быть использован в системах автоматической обработки естественного языка; потенциальные пользователи — это широкий круг российских и зарубежных исследователей и разработчиков-практиков, а также студентов. Ресурс будет распространяться по лицензии CC BY-SA, которая допускает модификацию и применение продукта как в некоммерческих, так и в коммерческих проектах.

Опыт использования тезаурусов других языков показывает, что у подобных ресурсов очень широкая область использования. Тезаурусы используются в большинстве современных систем автоматической обработки языка: в задачах классификации текстов, автоматического реферирования, генерации текстов, в информационном поиске, машинном переводе и др.

Таким образом, пользователями объединенного ресурса станет широкая аудитория исследователей в области фундаментальной и прикладной лингвистики, смежных гуманитарных наук, а также разработчики компьютерных систем и лексикографических продуктов в России и за рубежом. У тезауруса большой потенциал использования в рамках учебных курсов по компьютерной лингвистике, обработке естественного языка и т.п.

Благодарности

Исследование поддержано грантом РГНФ № 16-04-12019 «Интеграция тезаурусов RussNet и YARN». Исследование выполняется также при финансовой поддержке РФФИ в рамках научного проекта № 16-37-00354 мол_а «Методы автоматизации процесса коллективного построения лингвистических ресурсов».

Литература

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

[2] Лукашевич Н. В. Тезаурусы в задачах информационного поиска, М.: МГУ, 2011.

[3] Azarowa I.V. RussNet as a Computer Lexicon for Russian // Intelligent Information Systems 2008, ISBN 978-83-60434-44-4. P. 447–456.

[4] Сухоногов А.М., Яблонский С.А. Автоматизация построения англо-русского Wordnet // Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2005». М., 2005.

[5] Гельфейнбейн И.Г., Гончарук А.В., Лехельт В.П., Липатов А.А., Шило В.В. Автоматический перевод семантической сети WordNet на русский язык // Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2003». Протвино, Россия. 2003.

[6] Braslavski P., Ustalov D., Mukhin M., Kiselev Y. YARN: Spinning-in-Progress // Proceedings of the 8th Global Wordnet Conference, 2016.

[7] Ustalov D. Russian Thesauri as Linked Open Data // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference «Dialogue», 2015. P. 616–625.

RussNet and YARN thesauri integration

I. Azarova [1], P. Braslavski [2], V. Zakharov [1],
Y. Kiselev [2], D. Ustalov [3], M. Khokhlova [1]
[1] Saint-Petersburg State University
[2] Ural Federal University
[3] Krasovskii Institute of Mathematics and Mechanics

Thesauri and ontologies are widely used in many natural language processing tasks and applications. Wordnets are considered to be “a standard NLP tool” along with part-of-speech taggers, syntactic parsers, etc. The aim of the project is to integrate two large Russian-language lexicographic resources — RussNet and YARN (Yet Another RussNet).

Despite the fact that much has been done within RussNet and YARN projects, there is still no large open high-quality thesaurus for Russian to date. The best solution is to integrate both resources and develop them jointly in the future by representatives of these projects

Availability of a large open electronic thesaurus for Russian designed for natural language processing will substantially facilitate research and software production in such fields as artificial intelligence, information retrieval, computational linguistics in Russia, it will as well improve both effectiveness and performance of Russian language processing systems.

The problem of semantic resources' integration is being addressed within Linked Open Data (LOD) approach; concrete resources and language-specific features determine the originality of the methods can be applied to solve the problem.

We plan as a result of the project the integration of RussNet (40,000 entries, 30,000 synsets, 45,000 relationships) and YARN (about 120,000 entries, 46,000 synsets, and 30,000 hierarchical relationships) data, which will be cleared from inconsistencies, validated and completed by additional data. An important aspect of the project is a combination of crowdsourcing-based and expert-based approaches. Crowd management methodology is a new and relevant direction of research in many areas.

Keywords: lexical statistics, thesaurus, linguistic ontology, WordNet, dictionary integration.