УДК 004.912

ПОРОЖДЕНИЕ ТЕЗАУРУСА ТИПА WORDNET ДЛЯ РУССКОГО ЯЗЫКА¹

Лукашевич Н.В. (louk_nat@mail.ru)
МГУ им. М.В.Ломоносова, Москва, Россия

Лашевич Г. (design.berg@mail.com)
Казанский федеральный университет, Казань, Россия

Герасимова А.А. (anastasiagerasimova432@gmail.com)
МГУ им. М.В.Ломоносова, Москва, Россия

Иванов В.В. (nomemm@gmail.com), Казанский федеральный
университет, Казань, Россия

Добров Б.В. (dobrov_bv@mail.ru), МГУ им. М.В.Ломоносова,
Москва, Россия

В статье описывается автоматизированная процедура создания русского тезауруса типа WordNet (RuWordNet) путем трансформации тезауруса русского языка РуТез (опубликованной версии РуТез-lite 2.0). Было поставлено две цели: представление лексических отношений русского языка в виде сетей, соответствующих частям речи, и воспроизведение базового набора отношений WordNet. Первая версия RuWordNet создана и доступна для просмотра в Интернет и в виде xml файлов.

Ключевые слова: тезаурус, WordNet, автоматическая обработка текстов, лексические отношения

Введение

Лексические тезаурусы формата WordNet [Fellbaum, 1998], разработанного в Принстонском университете США (далее Принстонский WordNet), являются одним из наиболее востребованных ресурсов в

¹ Работа выполнена при финансовой поддержке РФНФ (проект № 15-04-12017).

различных задачах автоматической обработки текстов. Поэтому проекты создания таких ресурсов инициируются для многих языков.

Тезаурус WordNet представляет собой совокупность лексических понятий – синсетов (наборов синонимов), между которыми установлены такие отношения, как гипоним-гипероним (родовидовые отношения), мероним-холоним (отношения часть-целое) и некоторые другие

Известно по крайней мере четыре попытки создать ворднет для русского языка [Азарова и др., 2004; Гельфенбейн и др., 2003; Сухоногов и др. 2005; Браславский и др., 2012]. Последний проект YARN (Yet Another Russian wordNet) был инициирован в 2012 и создается на основе краудсорсинга. В настоящее время YARN содержит большое количество синсетов с незначительным количеством отношений между ними.

Для русского языка существует тезаурус РуТез, созданный для применения в задачах автоматической обработки текстов и информационного поиска [Лукашевич, 2011]. Структура и принципы создания РуТез отличаются от структуры WordNet. РуТез больше похож на онтологию, чем на лексический ресурс: понятия тезауруса имеют уникальные имена, отношения в тезаурусе представляют собой формализованные отношения между понятиями. Текущий размер опубликованной версии тезауруса РуТез (РуТез-lite 2.0), доступной для некоммерческого применения, составляет более 115 тысяч слов и выражений² [Loukachevitch et al., 2014]. РуТез может применяться во всех задачах, в которых обычно применяется WordNet, но исследователи и практики хотят иметь и для русского языка большой качественный ресурс, который можно было бы назвать русским ворднетом.

В данной статье описывается процесс трансформации данных тезауруса РуТез в структуры WordNet. При этом ставится задача воспроизвести две основных характеристики Принстонского WordNet, а именно, представление отношений между рядами синонимов в виде сетей, соответствующих отдельным частям речи, и базовый набор отношений. Получившийся ресурс RuWordNet также опубликован и может быть получен в виде xml-файлов для некоммерческого применения. Объем тезауруса RuWordNet также составляет 115 тысяч слов и выражений русского языка³.

1. Подходы к созданию тезаурусов типа WordNet

Создание больших лексических ресурсов типа WordNet с нуля представляет собой сложную задачу, которая требует усилий в течение многих лет [Азарова и др., 2004]. Чтобы ускорить разработку ворднета для

² <http://www.labinform.ru/pub/ruthes/index.htm>

³ <http://labinform.ru/pub/ruwordnet/index.htm>

своего языка, первая версия такого ресурса может быть создана посредством автоматического перевода Принстонского WordNet на целевой язык [Vossen, 1998; Гельфенбейн и др., 2003; Сухоногов и др. 2005], однако затем требуются значительные усилия на вычитку и правку полученного перевода.

В качестве промежуточного подхода исследователями предлагается двухэтапное создание ворднета своего языка: сначала перевод и перенос отношений верхних пяти тысяч понятий Принстонского WordNet (так называемый core WordNet), а затем ручное пополнение иерархий на основе словарей и корпусов своего языка. Такой подход был использован при создании таких ресурсов, как EuroWordnet [Vossen, 1998] и Датский ворднет – DanNet [Pedersen, 2010].

Проанализировав существующие подходы к развитию ворднетов своих языков, создатели финского ворднета (FiWN) решили вручную перевести Принстонский WordNet, используя труд профессиональных переводчиков. В результате финский ворднет был создан на основе перевода более чем 200 тысяч значений слов Принстонского WordNet в течение 100 дней.

В работе [Браславский и др., 2012] предлагается разрабатывать русский ворднет (YARN), используя русский Вики-словарь и краудсорсинг, т.е. привлекать для создания ворднета большое количества студентов и заинтересованных лиц.

2. Структура тезауруса WordNet

Структура Принстонского WordNet (и других ворднетов) строится на наборах синонимов – синсетов, которые организованы в иерархические сети для каждой части речи – существительных, прилагательных, глаголов и наречий. Синсеты каждой части речи имеют собственный набор отношений.

Наиболее часто используемым отношением между синсетами существительных является отношение гипоним - гипероним, которое устанавливается между родовыми и видовыми словами. С 2006 года из этих отношений были выделены отношения экземпляр-класс (Instance Hierarchy и Instance Hierarchy) [Miller et al., 2006].

Отношения между существительными включают также отношения часть-целое, которые подразделяются на собственно отношение часть-целое (*крыло – часть птицы*), элемент-группа (*дерево – элемент леса*) и материал (*снег – материал для снежков*).

Для всех частей речи может быть установлено отношений антонимии, которое является лексическим отношением, т.е. устанавливается между отдельными словами, а не синсетами. При этом отношение антонимии является основным отношением между качественными прилагательными (Gross, Miller, 1990), для которых описываются только отношения

антонимии и сходства (similarity). Например, для слова *heavy* (тяжелый) слово *light* (легкий) описано как антоним. Такие слова как *hefty*, *ponderous*, *massive* (объемный, тяжеловесный, массивный) связаны со словом *heavy* отношением сходства (similar to). Другие ворднеты, такие как GermaNet [Kunze et al., 2010] или Polish WordNet (PIWordNet) [Derwojedowa et al., 2008] изменили принципы описания прилагательных и используют отношения гипоним-гипероним для прилагательных.

Глаголы в WordNet в основном связаны отношения гипоним-гипероним, как и существительные. Кроме того, они имеют свой собственный набор отношений, которые не используются для описания существительных или прилагательных, например, отношение следствия (entailment: *buy – pay*) и отношение каузации (causation: *give – have, kill – die*). Отношение следствия устанавливается между двумя глаголами V_1 и V_2 , если из предложения "*Someone V₁*" логически следует "*Someone V₂*" и имеется временное включение события V_1 в событие V_2 . Отношение каузации также может быть классифицировано как подтип общего отношения логического следования, но в случае каузации нет временного включения между соответствующими ситуациями [Fellbaum, 1998].

3. Структура и отношения тезауруса PyTез

PyTез и WordNet – оба представляют собой тезаурусы, т.е. лексические ресурсы, в которых близкие по смыслу слова собраны в синсеты (WordNet) или понятия (PyTез), между которыми установлены отношения. При применении обоих тезаурусов к обработке текстов должны выполняться похожие этапы, включающие сопоставление текста с тезаурусом, использование при необходимости описанных отношений. При этом между тезаурусами есть и существенные различия.

Во-первых, в PyTез нет деления на лексические сети по частям речи; любые части речи могут быть связаны с одним и тем же понятием тезауруса PyTез, если они означают одно и то же (так называемые частеречные синонимы).

Во-вторых, в PyTез отношения устанавливаются только между понятиями. Основное отношение класс-подкласс приблизительно соответствует отношению гипоним-гипероним в WordNet.

Также в PyTез есть отношение часть-целое, однако в отличие от WordNet оно устанавливается только в тех случаях, когда часть всегда (или, по крайней мере, в подавляющем большинстве случаев) относится к указанному целому, т.е. не может быть указано множество альтернативных целых. Это позволяет с большей надежностью использовать транзитивность отношений часть-целое [Лукашевич, 2011].

Кроме того, одним из основных отношений в PyTез является отношение онтологической зависимости, которое показывает зависимость

существования одного понятия от другого. Примером такого отношения является отношение между понятиями ДЕРЕВО-ЛЕС, где ЛЕС является зависимым понятием, требующим существования понятия ДЕРЕВО. Таким образом, структура и набор отношений в тезаурусе РуТез значимым образом отличается от структуры и отношений WordNet.

4. Преобразование данных тезауруса РуТез в структуру тезауруса типа WordNet

Основными задачами при преобразовании данных тезауруса РуТез в ворднет русского языка (RuWordNet) были следующие:

- 1) разделить сеть понятий тезауруса на отдельные сети синонимов, соответствующие частям речи;
- 2) обеспечить набор отношений, которые соответствует тезаурусам типа WordNet.

Разделение на синсеты по частям речи было сделано на основе морфосинтаксического представления текстовых входов тезауруса РуТез, которое было создано автоматизированно. В результате текстовые входы тезауруса были разделены на синсеты трех частей речи: существительные (отдельные существительные и группы существительного), прилагательные (отдельные прилагательные и группы прилагательного), глаголы (отдельные глаголы и группы глаголов) (Табл. 1). Между разделенными синсетами, относящимися к одному исходному понятию тезауруса РуТез, были установлены отношения частеречной синонимии.

Табл. 1

Часть речи	Число синсетов	Число уник.входов	Число значений
Сущ.	29296	68695	77153
Глагол	7634	26356	35067
Прилаг.	12864	15191	18195

В RuWordNet отношения гипоним-гипероним установлены только между синсетами одной и той же части речи. Кроме отношений, перенесенных из РуТез, эти отношения включают в себя и отношения, полученные на основе использования свойства транзитивности: если исходное понятие тезауруса РуТез не было связано с какой-то частью речи, а его нижестоящие и вышестоящие понятия имели текстовые входы этой части речи, то отношения гипоним-гипероним устанавливалось непосредственно между вышестоящими и нижестоящими синсетами в RuWordNet.

Также как и в последних версиях WordNet, в RuWordNet установлены отношения экземпляр-класс. В настоящее время эти отношения

установлены только между синсетами географических объектов и синсетами их типов (*Москва – столица*).

Отношения часть-целое были автоматизированно выгружены из RuГез и скорректированы в соответствии с традициями ведения ворднетов. Таким образом, отношения часть-целое в RuWordNet включают такие подтипы как функциональные части (*ноздри – нос*), ингредиенты (*включение – вещество*), географические части (*Севилья – Андалузия*) и др.

Прилагательные в RuWordNet, подобно немецкому и польскому ворднетам, также соединены отношениями гипоним-гипероним, например, прилагательное *цветовой* является гиперонимом для прилагательных *красный, синий, зеленый* и др.

Текущий набор отношений между глагольными синсетами, помимо отношений гипоним-гипероним, включает также отношения часть-целое. Например, синсет {*видеть во сне, сниться, ...* } описан как часть для синсета {*спать, поспать..* }. Такой отношение для соответствующих глагольных синсетов [*to dream – to sleep*] существует и в Принстонском WordNet и называется отношением следования. Как указывалось в [Fellbaum, 1998], отношения следования между синсетами глаголов напоминают отношения часть-целое между синсетами существительных.

Отношения антонимии в настоящее время представлены как концептуальные отношения в RuWordNet, т.е. они устанавливаются между синсетами, а не единичными лексемами. Они введены для всех частей речи, в основном для синсетов, которые описывают свойства и состояния, например, синсет {*легкость, с легкостью, без труда, без затруднений*} связан отношением антонимии с синсетом {*тяжесть, трудность*}.

Текущее число отношений по типам и частям речи в текущей версии RuWordNet представлено в Табл. 2.

Табл.2

Часть речи	Гиперонимы	Экземпляры	Часть-Целое	Частеречн. синонимы	Антонимы
Сущ.	39155	1863	10010	18179	455
Глагол	10440	0	117	7451	20
Прилаг.	17834	66	829	14139	457

Чтобы сравнить представление лексики русского языка в тезаурусах RuГез и RuWordNet, рассмотрим описание слова *двор*. В тезаурусе RuГез (Рис. 1), просматривая слово *двор*, можно увидеть, что это слово соответствует трем понятиям, каждое из которых имеет однозначное имя. Текстовыми входами понятий, помимо слова *двор* являются не только существительные, но и прилагательные.

В RuWordNet (Рис. 2) значения представлены соответствующими синсетами, прилагательные выделены в отдельные синсеты, к которым

ведет отношение «частеречный синоним». В целом, RuТез выглядит более похожим на онтологию, а RuWordNet на лексическую сеть.

Текстовый вход: ДВОР

УСАДЬБА

(ДВОР, ПОДВОРЬЕ, УСАДЕБНЫЙ, УСАДЬБА)

ВЫШЕ [ЗАГОРОДНЫЙ ДОМ](#)
ВЫШЕ [ИНДИВИДУАЛЬНЫЙ ЖИЛОЙ ДОМ](#)
ЦЕЛОЕ [ПОМЕСТЬЕ](#)
НИЖЕ [ХУТОР](#)
ЧАСТЬ [ПРИУСАДЕБНЫЙ УЧАСТОК](#)

ДВОР (ЗЕМЕЛЬНЫЙ УЧАСТОК)

(ДВОР, ДВОРИК, ДВОРОВОЙ, ДВОРОВОЙ УЧАСТОК, ПРИДОМОВЫЙ УЧАСТОК)

ВЫШЕ [ЗЕМЕЛЬНЫЙ УЧАСТОК](#)
НИЖЕ [ГОРОДСКОЙ ДВОР](#)
НИЖЕ [ПРИУСАДЕБНЫЙ УЧАСТОК](#)

КОРОЛЕВСКИЙ ДВОР

(ДВОР, КОРОЛЕВСКИЙ ДВОР, ЦАРСКИЙ ДВОР)

ВЫШЕ [ОКРУЖЕНИЕ, ОКРУЖАЮЩИЕ ЛЮДИ](#)

Рис.1. Представление значений слова *двор* в тезаурусе RuТез

Текстовый вход: ДВОР

Синсет: [ДВОР](#) [УСАДЬБА](#) [ПОДВОРЬЕ](#)

ГИПЕРОНИМ [ДОМ ДЛЯ ОДНОЙ СЕМЬИ](#) [ДОМУСАДЕБНОГО ТИПА](#) [ОДНОСЕМЕЙНЫЙ ДОМ](#) [ИНДИВИДУАЛЬНОЕ ЖИЛЬЕ](#) [ИНДИВИДУАЛЬНЫЙ ЖИЛОЙ ДОМ](#) [ИНДИВИДУАЛЬНЫЙ ЖИЛИЩНЫЙ ФОНД](#) [ИНДИВИДУАЛЬНОЕ ДОМОВЛАДЕНИЕ](#)

ГИПЕРОНИМ [ЗАГОРОДНЫЙ ДОМ](#)

ГИПОНИМ [ХУТОР](#) [ХУТОРОК](#)

ЦЕЛОЕ [ИМЕНИЕ](#) [ПОМЕСТЬЕ](#)

ЧАСТЕРЕЧНЫЙ СИНОНИМ [УСАДЕБНЫЙ](#)

ЧАСТЬ [ПОДВОРЬЕ](#) [ПРИУСАДЕБНАЯ ЗЕМЛЯ](#) [ПРИУСАДЕБНЫЙ УЧАСТОК](#) [ПРИУСАДЕБНЫЙ ЗЕМЕЛЬНЫЙ УЧАСТОК](#)

Синсет: [ДВОР](#) [ЦАРСКИЙ ДВОР](#) [КОРОЛЕВСКИЙ ДВОР](#)

ГИПЕРОНИМ [КРУГ СРЕДА](#) [БЛИЗКОЕ ОКРУЖЕНИЕ](#) [ОКРУЖЕНИЕ БЛИЖАЙШЕЕ](#) [ОКРУЖЕНИЕ](#) [ОКРУЖАЮЩИЕ ЛЮДИ](#)

Синсет: [ДВОР](#) [ДВОРИК](#) [ДВОРОВОЙ УЧАСТОК](#) [ПРИДОМОВЫЙ УЧАСТОК](#)

ГИПЕРОНИМ [ЗЕМЛЯ НАДЕЛ](#) [НАДЕЛ ЗЕМЛИ](#) [ДЕЛЯНКА](#) [ЗЕМЛИЦА](#) [УЧАСТОК СУЩИ](#) [ДЕЛЯНКА ЗЕМЛИ](#) [УЧАСТОК ЗЕМЛИ](#) [ЗЕМЕЛЬНЫЙ НАДЕЛ](#) [ЗЕМЕЛЬНЫЙ УЧАСТОК](#)

ГИПОНИМ [ПОДВОРЬЕ](#) [ПРИУСАДЕБНАЯ ЗЕМЛЯ](#) [ПРИУСАДЕБНЫЙ УЧАСТОК](#) [ПРИУСАДЕБНЫЙ ЗЕМЕЛЬНЫЙ УЧАСТОК](#)

ГИПОНИМ [ДВОРОВАЯ ТЕРРИТОРИЯ](#) [ГОРОДСКОЙ ДВОР](#) [ПРИДОМОВАЯ ТЕРРИТОРИЯ](#)

ГИПОНИМ [ЗАДНИЙ ДВОР](#)

File:///D:/RUPTES/Output/WordNet/2016/ruwordnet/Output/Нoun/Нer/33000/00101.htm

Рис.2. Представление значений слова *двор* в тезаурусе RuWordNet

Заключение

В статье описан автоматизированный процесс трансформации тезауруса русского языка RuТез (в версии RuТез-lite 2.0) в тезаурус типа WordNet. Полученный тезаурус получил имя RuWordNet. Для этого были

воспроизведены две основные особенности тезаурусов типа WordNet: разделение сети понятий на лексические сети для каждой части речи, а также воспроизведение основного набора отношений.

Оба тезауруса, RuТез-lite 2.0 и RuWordNet, опубликованы как статические веб-страницы. Также RuWordNet может быть доступен через веб-интерфейс⁴. Исследователи могут получить оба типа тезаурусов в виде xml-файлов и сравнить их в конкретных приложениях.

Список литературы

- [Азарова и др., 2004] Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог-2004. 2004. С. 542-547.
- [Браславский и др., 2012] Браславский П., Мухин М., Ляшевская О.Н., Бонч-Осмоловская А.А., Кржижановский А., Егоров П., Власенко С.В. YARN: начало // Вестник Тверского государственного университета. Серия: Филология, 10(2). 2012. С.13-20.
- [Гельфейнбейн и др., 2003] 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 Диалог по компьютерной лингвистике и её приложениям, Протвино, Россия. 2003.
- [Лукашевич 2011] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Изд-во Московского университета, 2011.
- [Сухоногов 2005] Сухоногов А.М., Яблонский С.А. Автоматизация построения англо-русского WordNet // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог-2005, 2005.
- [Derwojedowa et al., 2008] Derwojedowa M., Piasecki M., Szpakowicz S., Zawisławska M., Broda B. Words, concepts and relations in the construction of Polish WordNet // In Proc. of the Global WordNet Conference, 2008. pp. 162-177.
- [Fellbaum 1998] Fellbaum C. A semantic network of English verbs // WordNet: An electronic lexical database. 1998. pp.153-178.
- [Gross et al., 1990] Gross D., Miller K.J. Adjectives in WordNet // International Journal of Lexicography, 3(4), 1990. pp.265-277.
- [Kunze et al., 2010] Kunze C., Lemnitzer L. Lexical-Semantic and Conceptual relations in GermaNet // In Storjohann P (ed) Lexical-semantic relations: Theoretical and practical perspectives, 2010. pp. 163-183.
- [Lindén et al., 2014] Lindén K., Niemi J. Is it possible to create a very large wordnet in 100 days? An evaluation // Language resources and evaluation, 48.2, 2014. pp. 191-201.
- [Loukachevitch et al., 2014] Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I. RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes

⁴ <http://ruwordnet.ru/>

- // In proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2014. 2014. pp. 340-350.
- [Miller 1998] Miller G. Nouns in WordNet // In WordNet: An Electronic Lexical Database, Fellbaum, C (ed), The MIT Press, 1998. pp. 23-47.
- [Miller et al., 2006] Miller G.A., Hristea F. WordNet nouns: Classes and instances // Computational linguistics, 32(1), 2006. pp.1-3.
- [Pedersen et al., 2009] Pedersen B.S., Nimb S., Asmussen J., Sørensen N.H., Trap-Jensen L.,d Lorentzen, H. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary // Language resources and evaluation, 43(3) 2009. pp.269-299.
- [Vossen 1998] Vossen P. Introduction to EuroWordNet // EuroWordNet: A multilingual database with lexical semantic networks. Springer Netherlands, 1998, pp. 1-17.

GENERATING RUSSIAN WORDMET

N. Loukachevitch (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

G. Lashevich (design.berg@mail.com),

Kazan Federal University, Kazan, Russia

A.Gerasimova (anastasiagerasimova432@gmail.com)

Lomonosov Moscow State University, Moscow, Russia

V. Ivanov (nomemm@gmail.com)

Kazan Federal University, Kazan, Russia

B. Dobrov (dobrov_bv@mail.ru),

Lomonosov Moscow State University, Moscow, Russia

. In this paper we have described the semi-automatic process of transforming the Russian language thesaurus RuThes (in version, RuThes-lite 2.0) to WordNet-like thesaurus, called RuWordNet. In this procedure we attempted to achieve two main characteristic features of wordnet-like resources: division of data into part-of-speech-oriented structures with cross-references between them and providing a set of relations similar to WordNet-like resources.

Keywords: thesaurus, WordNet, natural language processing, lexical relations