Разработка компьютерного тезауруса русского языка типа WordNet

Азарова И.В., Митрофанова О.А., Синопальникова А.А., Ушакова А.А., Яворская М.В.

Введение: история и эволюция WordNet

В связи с необходимостью обеспечивать представление семантической информации для текстов на разных языках в настоящее время широко используются системы тезаурусного представления лексики типа WordNet. На данный момент для различных языков уже разработан целый ряд подобных представлений лексических систем.

Пионерами в этой области были ученые Принстонского Университета (США) во главе с Дж. Миллером, начавшие в 1985 году работу над созданием компьютерного идеографического словаря, который охватывал бы всю лексику английского языка. Первоначально задачей компьютерного представления было моделирование процессов использования семантической информации человеком, поэтому при построении американские ученые активно использовали различные психолингвистические методики. Результатом их работы стала лексико-семантическая база данных **WordNet**.

Данный словарь не содержит сведений о произношении, этимологии или рекомендаций по использованию слов в текстах: в нем хранится базовая информация о соотнесении слов-означающих и означаемых. Принстонский WordNet состоит из четырех отдельных лексических баз данных: блока существительных, блока глаголов, блока прилагательных и блока наречий. Основной конструктивной единицей в словаре WordNet является синонимический ряд (синсет) со стандартным определением словарного типа. Предполагается, что синсет в словаре представляет лексикализованное понятие. Если слово появляется более, чем в одном синсете, то оно считается многозначным (при этом омонимия рассматривается как разновидность многозначности).

Слова и синсеты связаны друг с другом определенными парадигматическими отношениями. Синонимия отражается в объединении слов в класс эквивалентности — синсет. На синсетах задаются парадигматические отношения: антонимические, гипонимические, меронимические и различные виды лексического вывода - каузация, пресуппозиция (о различных типах отношений мы поговорим позже).

Запрос к тезаурусу задается в виде слова (словоформы). При запросе выводится следующая информация:

- к какой части речи может относится данное слово (словоформа);
- список всех синонимических рядов (**синсетов**), в которые входит слово (в рамках определенной части речи);

- для каждого синсета: приводятся **примеры** употребления слова в этом значении; дается словарное **определение (его значения)** стандартного типа; указываются все его парадигматические связи: **гиперонимы, гипонимы, меронимы, холонимы, тропонимы** и пр.

Английская версия WordNet в 1999 г. содержала около 122 тысяч слов-означающих, организованных в 100 тысячах синсетов, на которых задано около 139 тысяч отношений.

Важно осознание того, что WordNet создавался в рамках определенной лингвистической теории, поэтому в оригинальной версии словаря не представлены тематически организованные поля (то есть внеязыковая картина мира). Нельзя извлечь, например, все слова, которые относятся к такой тематической области, как игра в футбол: данные слова будут разбросаны по всему словарю и никак не связаны между собой. С точки зрения Дж. Миллера, тематическая организация лексики — другой тип структуры, отличающийся от той структуры, которая задается в WordNet. В частности, оригинальный вариант словаря не содержал информации о том, как распознать различные значения слов, то есть в нем не было указаний на контекстуальные маркеры, которые позволили бы принять решения об отнесении значения к тому или иному синсету.

Еще одним важным аспектом WordNet является представление лексических баз части речи не в виде единой структуры (дерева с абстрактным корнем типа "нечто"), а виде набора иерархических структур (леса). Весьма интересен в данном отношении набор вершин-корней. Для существительных задано 25 семантических областей: ВЕЩЕСТВО, ВЛАДЕНИЕ, ВРЕМЯ, ДЕЙСТВИЕ (ДЕЯТЕЛЬНОСТЬ), ЖИВОТНЫЕ (ФАУНА), ЗНАНИЕ, КОЛИЧЕСТВО, ЛИЦО, МЕСТОПОЛОЖЕНИЕ, НАМЕРЕНИЕ, ОБЩЕНИЕ, ОБЪЕКТ, ОТНОШЕНИЕ, ПИЩА, РАСТЕНИЕ, СОВОКУПНОСТЬ, СОБЫТИЕ, СОЗДАНИЕ ЧЕЛОВЕКА, СОСТОЯНИЕ, ТЕЛО, ФОРМА, ЧЕРТА (СВОЙСТВО), ЧУВСТВО, ЯВЛЕНИЕ. Для глаголов выделено 38 основных семантические областей: СОБЫТИЕ, СОСТОЯНИЕ, ДЕЙСТВИЕ, ПУТЬ, ОБРАЗ ДЕЙСТВИЯ, МЕСТО, ОТРИЦАНИЕ и т.д.

Принстонский **WordNet** распространяется свободно.

С 1996 по 1999 гг. в рамках проекта "**EuroWordNet**" было создано уже несколько подобных систем для европейских языков (французского, немецкого, испанского, итальянского, голландского, эстонского и чешского). Все они были построены на базе американского WordNet'a (по единой модели), хотя могли содержать особые парадигматические отношения. Каждый из национальных тезаурусов WordNet отражает все особенности лексической системы соответствующего языка, сходство между языками выражается (неявно) в сходстве структур. Явным образом национальные тезаурусы были связаны

при помощи общей понятийной схемы (Top Ontology), которая вначале включала 63 общих понятия для существительных и глаголов, а затем была расширенна за счет предложений участников проекта до 1360. Элементы этой общей понятийной схемы связаны между собой при помощи системы межъязыковых индексов (**Inter-Lingual-Index**) в большую многоязычную базу данных. Посредством данного индекса можно переходить от одной языковой структуры WordNet к другой. Лексические противопоставления, не входящие в обобщенный онтологический базис, сохраняются в конкретных языковых реализациях WordNet. Таким образом, база EuroWordNet может использоваться для информационного поиска не только в рамках одного языка, но и для многоязычного поиска.

Основные отличия между EuroWordNet и WordNet заключаются в следующем:

1. При построении WN 1.5. американские ученые опирались, в основном, на **психолинвистические** данные, тогда как для построения национальных компонентов EuroWordNet использовались уже традиционные **лексикографические** источники (электронные и бумажные словари: толковые, переводные и синонимические) и корпусы текстов;
2. В структуру EuroWordNet были включены отношения, позволяющие фиксировать связи между словами **разных частей речи**, например, Near_Synonymy - для представления таких связей, как *to endorse - endorsement, to run - run*, Role_Relations - для отображения синтагматических связей глаголов и существительных таких, *как to hammer - carpenter, to play - musician*.

Хотя проект EuroWordNet был завершен в 1999 году, работы по созданию ресурсов типа WordNet для новых языков (норвежского, датского, шведского, португальского, греческого, румынского, литовского, болгарского, польского, венгерского, турецкого, арабского, хинди, корейского, китайского и др.) продолжаются, и в случае совместимости вновь созданных ресурсов они подключаются к общей системе.

В 2001 г. с целью объединения уже существующих и только развивающихся национальных WordNet'ов была создана **Всемирная Ассоциация WordNet**.

RussNet: основные понятия, структура

На кафедре математической лингвистики СПбГУ вот уже 3 года ведется работа по созданию лексикона типа WordNet для русского языка. RussNet унаследовал основные черты Принстонского WN и EuroWN.

1. Сохранена общая установка на отображение лексической системы языка в **целом** (а не только узкоспециальной, терминологической лексики), причем это должен быть, с

одной стороны, словарь базовой лексики, активно употребляемой на протяжении длительного периода развития языка, с другой строны, словарь активной лексики широко употребляемой в газетно-публицистическом жанре в последние десятилетия.

2. Как и каждый из компонентов EuroWN, RussNet состоит из **4 взаимосвязанных баз данных** для основных частей речи: существительных, глаголов, прилагательных и наречий.

3. Основными структурными единицами RussNet являются **слово** и **синонимический ряд (синсет)**.

4. Слова, связанные между собой отношением **синонимии,** образуют базовые единицы словаря — синсеты.

5. На синсетах заданы синтагматические и парадигматические отношения **гипонимии/гиперонимии**, тропонимии, меронимии/холонимии, отношения каузации, пресуппозиции, лексического вывода и пр., для отдельных ЛСВ слов — отношение **антонимии;** для отдельных слов — **деривационные** отношения.

Синонимия и синонимические ряды

Синсет как лексикализованное понятие объединяет в себе все лексико-семантические варианты слов, выражающие данное понятие, связанные между собой отношением синонимии. Существует по крайней мере 2 разных подхода к определению синонимии:

- **взаимозаменяемость в контексте**: два слова считаются синонимами, если существует высказывание (или ряд высказываний), в котором замена этих слов друг на друга не влияет на истинность высказывания (Лайонз 1977, Миллер 1990, Апресян 1991);
- **семантическая близость слов**: в данном случае в качестве критерия рассматривается наличие у слов некоего общего значения (Евгеньева 1977).

Применение этих критериев на практике имеет свои плюсы и минусы: с одной стороны, в языке существует не так много абсолютных синонимов (слов, взаимозаменяемых в любом контексте, таких как *гиппопотам - бегемот*), с другой стороны, критерий семантической близости достаточно субъективен: трудно определить без количественного ананлиза значений, насколько "незначительны" семантические различия между синонимами.

Для удобства пользователя каждый синсет дополняется словарным определением и примерами употребления слов в контексте.

Родо-видовые отношения

Среди семантических отношений в RussNet особую роль играют родо-видовые отношения (гипонимия/гиперонимия и тропонимия), позволяющие организовывать синсеты в деревья зависимостей, выстраивать их иерархию. (В рамках гипонимических структур происходит наследование свойств, присущих гипероним у.)

Степень родо-видовой иерархии слов для разных частей речи различна. **Существительные** — дают типичную родовидовую иерархию, которая тем не менее не обладает слишком большой глубиной (не больше 5-8 уровней). Для **глаголов** в качестве родовидового отношения выступает нечто иное, это отношение даже называется иначе — **тропонимия** (от греческого *τροποσ* — способ). Характерным примером глагольных родовидовых отношений является пара *идти — хромать*. Для **прилагательных** и **наречий** родовидовые отношения либо вообще не представлены, либо представлены без явной системы, лишь для некоторых групп, что заставляет нас искать другие способы структурирования синсетов.

Верхние ярусы лексической базы RussNet для существительных и глаголов будут согласованы с общей понятийной схемой **(Top Ontology)**; мы намерены завершить координацию в течение этого года.

Меронимия, каузация, лексический вывод, конверсия

Типичным для тезаурусных представлений является отношение **меронимии** ("часть - целое"). Обычно в качестве его основания рассматривают логические связи между понятиями типа "компонент-предмет" (*ветка-дерево*), "член-множество" (*дерево-лес*), "материал-предмет" (*алюминий-самолет*), хотя можно представить и другие возможные связи, которые можно классифицировать как меронимические, например: "порция-масса" (*кусок-пирог*) или "место-область" (*Москва-Россия*). Однако ясно, что при анализе материала на очень низком уровне (атомов), вряд ли можно сказать, что атом — часть конкретного предмета. Соотнесение частей с целым заканчивается там, где уменьшение частей перестает служить целям разграничения сходных предметов. Меронимическая иерархия не похожа на дерево, скорее на сеть: *острие — часть ножа, иглы, карандаша, булавки* и т.п.

Менее типичными для тезаурусных представлений являются такие отношения в глагольной лексике, как **каузация** (каузативный глагол — результирующее состояние каузации, например, *убить - умереть*, *высушить - стать сухим* и т. п.); отношение сложного действия и его части, оно было названо отношением **лексического вывода** (например, *спать - храпеть, красить - мазать* и т.п.); отношение **пресуппозиции** (дей-

ствие -- необходимое предыдущее действие, например, *выиграть - играть, развязать - завязать*).

В структуре RussNet мы планируем представлять также **конверсию** (в английском варианте она рассматривается как антонимия).

Дополнительно нами были внесены некоторые изменения в структуру компьютерного тезауруса, позволяющие представлять информацию о важных отношениях между лексическими единицами русского языка.

Деривационные отношения

Мы планирует указывать **деривационные словообразовательные** отношения, разделяя следующие подтипы.

a. **Транспозиция.** Например, *читать - чтение, белый - белизна*: выражается одно и то же понятие, но меняется категориальное значение слова.

b. **Экспрессивная синонимия**. Например, между словами *дом* и *домик, белый* и *беленький* существует определенная семантическая близость, но при этом они не могут быть взаимозаменяемы в контексте, что не позволяет считать их синонимами в полном смысле этого слова. Модели экспрессивного словообразования отличаются большим разнообразием и продуктивностью, в русском языке они активно используются, такие отношения необходимо так или иначе фиксировать. Включать экспрессивные синонимы в синсеты напрямую представляется нам некорректным, поэтому мы вводим новый тип отношений "экспрессивная синонимия". Экспрессивные синонимы указываются как расширение соответствующего синсета.

Синтагматические отношения

В рамках EuroWN помимо парадигматических отношений представлены и различные **синтагматические** отношения, связывающие слова, которые принадлежат различным частям речи: прилагательные и существительные, глаголы и существительные, глаголы и прилагательные, и т. п., что дает нам возможность использовать WordNet и для **контекстного поиска**:

a. для ЛСВ глагола указываются грамматические структура валентностей;

b. для ЛСВ прилагательных указываются классы существительных, сочетаемость с которыми можно предсказать, исходя из значения прилагательных.

Методы и источники, используемые при построении RussNet

Необходимость представить в рамках RussNet разнообразную информацию о семантических связях слов в системе русского языка:

- парадигматических (системных связей слов в лексико-семантических группах),
- деривационных (семантической организации словообразовательных гнезд) и
- синтагматических связей (взаимодействия значений слов на уровне текста),

требует использования разнообразных **источников** информации о содержании слова.

В их число входят:

- лексикографические данные (толковые, синонимичиские и словообразовательные словари);
- употребление изучаемых слов в текстах (использовались корпусы газетных и литературных текстов, относящихся к концу 20 - началу 21 века);
- данные, полученные в психолингвистических экспериментах (ассоциативные словари).

Разнообразие источников обуславливает комплексный характер методов, применяемых для извлечения информации о семантических отношениях внутри лексической системы.

На этапе **дефиниционного анализа** проводится качественный и количественный состав семантических признаков, существенных для конкретной лексико-семантической группы. Определяются семантические отношения внутри отдельных ЛСГ. При дефиниционном анализе использовались словари средних объемов. По-видимому, в наибольшей степени целям семантического исследования из толковых словарей русского языка соответствует МАС.

Обращение к **деривационному анализу** необходимо тогда, когда в лексико-семантической группе обнаруживается разветвленная система словообразовательных связей, поскольку в подобных случаях процедуру выделения сем, выполняемую только на основании словарных дефиниций, нельзя считать достаточной. Важность введения в семантическое описание еще одного измерения, деривационного (или эпидигматического), становится очевидной при изучении лексико-семантических групп русского языка с высоким числом производных, где необходимо учитывать связи семантических признаков и словообразовательных морфем в значении производных.

Значение слова определяется не только его местом в лексико-семантической группе или в семантическом поле, но также и потенциально возможными сочетаниями с другими словами, которые составляют лингвистический контекст. На следующем этапе иссле-

дования — на этапе **контекстного анализа** — осуществляется детальное изучение семной структуры отдельных значений каждого из слов, входящих в состав той или иной лексико-семантической группы. Набор типовых контекстов для лексической единицы, учитывающий модели лексической и синтаксической сочетаемости данного слова, определяет все множество значений, свойственных лексеме. Процедура контекстного анализа сводится к следующему. Для каждого вхождения исследуемых слов в текст в его окружении выявлялись те элементы, которые служат индикаторами реализации того или иного варианта семантического признака в значении слова, то есть элементы, составляющие уточняющий контекст. При этом учитывались средства морфемного, морфологического, лексического и синтаксического уровней; рассматривался как микроконтекст — непосредственное синтаксическое окружение лексической единицы, в котором она способна реализовать свое значение, включаясь в общий смысл фрагмента текста, так и макроконтекст — окружение лексической единицы, не ограниченное синтаксическими зависимостями данного слова и, возможно, выходящее за пределы предложения, то окружение, которое позволяет определить функцию слова в организации текста. В итоге каждому вхождению изучаемых слов в текст ставится в соответствие набор сем, описывающий реализуемое в контексте значение.

Возможности практического применения RussNet

Популярность WordNet'a и его широкое распространение обусловлены тем, что подобный компьютерный тезаурус, являясь одновременно и справочной системой, и инструментом для проведения различных исследований, предоставляет пользователю целый набор ранее недоступных возможностей. WordNet породил качественно новые сферы исследований и стимулировал активное развитие уже существующих, притом не только в лингвистике, но и в педагогике, социологии, психологии, компьютерных технологиях.

WordNet как одноязычный лексический ресурс

- Наиболее активно WordNet используется в области **информационного поиска** .
    1. Данные WordNet удобно использовать для эффективного изменения параметров запроса пользователя:
        a. за счет **парадигматических** связей

Например, пользователь имеет возможность включить в запрос все компоненты синсета, куда входит запрошенное слово в интересующем пользователя значении, вместе с его гипонимами и согипонимами, исключив при этом все слова из других синсетов, которые содержат данное слово.

b. за счет **синтагматических** связей

Представление в WordNet связей слов дает также возможность осуществлять **контекстный поиск.**

2. В данной области информация WordNet используется также и для решения классической задачи снятия неоднозначности смысла слова (**Word Sense Disambiguition**). Главное преимущество WordNet состоит в том, что поскольку главную роль при решении неоднозначности играет контекст рассматриваемого слова, то подобный тезаурус предоставляет во всем объеме информацию о других словах (понятиях), связанных с данным, которые в этом случае должны или не должны появляться в контексте.
3. Кроме того, WordNet может использоваться в системах информационного поиска как средство **измерения смысловой близости текстов (Similarity Measurement)**, позволяющее уменьшать количества "мусора", выдаваемого при осуществлении поисковых запросов на конкретную тему. На кафедре математической лингвистики проводился ряд исследований, посвященных разработке алгоритма для вычисления смысловой близости текстов на основе гиперонимических отношений, содержащихся в WordNet.

- **Автоматическая частеречная разметка текста** (Part-of-Speech Tagging): опираясь на семантические связи слов, представленные в WordNet, мы можем определять их частеречную принадлежность.
- WordNet можно использовать и как лексикон для **формальных грамматик** (например, AGFL).
- WordNet является удобным формализмом для представления, фиксирования лексического наполнения и отношений в лексике **подъязыков** (например, медицинских, экономических терминов) и многие учреждения в данный момент создают необходимые для своих целей WordNet-представления специальной лексики.
- При **изучении языков** WordNet часто используется как эффективное средство для быстрого изучения лексики и раскрытия содержания слов. При этом, в зависимости

от уровня подготовки изучающего, может варьироваться как количество выдаваемой ему по слову информации (сам синсет и связанные с ним другие синсеты), так и ее качество(синсеты, связанные с данным различными отношениями).

WordNet как многоязычный лексический ресурс

- При помощи EuroWordNet мы можем сравнивать WordNet-ы для разных языков: можно извлечь информацию об особенностях лексического наполнения языка, общей структуры связей между лексическими единицами и многое другое. Удобство использования WordNet в области межязыковых соответствий обеспечивается тем, что лексика всех языков, представленных в EuroWordNet объединена при помощи системы межъязыковых индексов (ILI) в одну общую структуру.
- Использовать WordNet для **автоматического перевода** текстов напрямую не представляется разумным. Межъязыковые индексы связывают семантически родственные слова, что в принципе позволяет извлекать так называемые переводные эквиваленты. Но в этой области WordNet можно использовать только как вспомогательное средство, и только в сочетании с системами синтаксического и морфологического анализа.
- Существование межъязыковых индексов позволяет успешно осуществлять **поиск** необходимой информации на нескольких языках.

Состояние RussNet на сегодняшний день

С помощью частотных словарей и корпусов текстов нами был проведен отбор наиболее частотной и значимой лексики современого русского языка.

На данный момент нами обработано более 2500 ЛСВ (около 1000 существительных, 1000 глаголов, 500 прилагательных), которые организованы приблизительно в 1000 синсетов.

Построение родо-видовых деревьев для русского языка проводится сверху вниз: на первом этапе исследуется определенное количество лексико-семантических групп слов, выявляется структура отношений внутри каждой из групп. Затем синонимы организуются и в синсеты, для каждой части речи строится набор гипонимических деревьев. На следующем этапе фиксируются остальные семантические отношения между синсетами: меронимия, каузация, различные деривациооные отношения. Синсеты, находящиеся на верхних уровнях иерархии включаются в понятийную схему (Top Ontology) RussNet. Мы

не копируем EuroWordNet Top Ontology, но планируем скоординировать полученную нами понятийную схему с ней.

Пока все данные RussNet хранятся в нескольких разрозненных файлах. На данный момент группа студентов занимается разработкой единого XML формата для представления данных, который позволил бы связать RussNet в единую лексическую базу. При этом, мы активно используем опыт, накопленный нашими немецкими коллегами при построении WordNet'a для немецкого языка (GermaNet).

Литература.

Fellbaum C. WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA,1998.

Lyons J. Semantics. (2 vol.) London and New York, 1977.

Miller G. et al. Five Papers on WordNet. CSL-Report, vol.43. Princeton University, 1990. ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps

Naumann, K. GermaNet. 2000. http://www.sfs.nphil.uni-tuebingen.de/Adj.html

Pala K., Sevecek P. The Czech WordNet, EuroWordNet (LE-8928). Deliverable 2D014, 1999. http://www.hum.uva.nl./~ewn/docs.html

Vossen, P. EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht: Kluwer, 1998.

Апресян Ю.Д. Новый объяснительный словарь синонимов русского языка. М., 1997.

Караулов Ю. Н., Уфимцева А. А. и др. Русский ассоциативный словарь. В 4-х томах. М, 1994.

Леонтьев А. А., Клименко А. П., Супрун А. Е. и др. Словарь ассоциативных норм русского языка./ Под ред. А. А. Леонтьева. М., 1977.

Митрофанова О. А. Структурный анализ сигнификативного значения: на материале глаголов процесса мышления английского и русского языков. Дисс. ….канд. фил. наук. СПб, 1999.

Объяснительный словарь русских глаголов./ под ред. Л.Г. Бабенко. М., 1999.

Ожегов С. И. Словарь русского языка. М., 1984.

Ожегов С. И., Шведова Н. Ю. Толковый словарь русского языка. М., 1992.

Словарь синонимов русского языка./ под ред А. П. Евгеньева. Л., 1970.

Словарь современного литературного русского языка. (т. 1-17). М.-Л., 1991.