ЦИФРОВАЯ КУЛЬТУРА

УНИВЕРСИТЕТ ИТМО

# Лексико-семантические ресурсы в автоматической обработке текстов

Высшая школа цифровой культуры

Университет ИТМО

dc@itmo.ru

# Оглавление


**Введение** 3

**Лексические отношения** 3

**Тезаурусы типа WordNet** 6

**Вычисление семантической близости текстов на основе тезауруса** 9

**Подходы к созданию тезаурусов типа WordNet для других языков** 11

**Тезаурусы типа WordNet для русского языка** 13

**Википедия как многоязычный онтологический ресурс** 15

**Заключение** 18

**Литература к лекции** 19

# Введение

В настоящее время большое значение в приложениях автоматической обработки текстов имеют ресурсы, описывающие знания о языке и мире в форме семантических сетей, т.е. понятий и отношений между ними. Такие ресурсы должны быть большого объема, чтобы их можно было эффективно использовать в прикладных задачах.

Одним из известных ресурсов такого рода является тезаурус WordNet, созданный для английского языка. В WordNet описаны лексические отношения между более чем для 150 тысяч слов и выражений. Знание лексических отношений полезно для таких приложений как:

- Распознавание перифраз – предложений с одним и тем же смыслом,
- Определение связности между предложениями в тексте,
- Расширение запроса при информационном поиске,
- Вопросно-ответные системы и др.

Такие лексико-семантические ресурсы можно использовать в комбинированных подходах в сочетании со статистическими методами, методами машинного обучения, поскольку они содержат готовые к использованию знания.

Терминологические ресурсы в форме семантических сетей (тезаурусы, онтологии, графы знаний) также востребованы в различных предметных областях для представления знаний о предметной области. Особенно большое внимание таким ресурсам уделяется в биомедицинских областях.

В данной лекции будут рассмотрены типы лексических отношений, которые востребованы в автоматической обработке текстов, проблемы представления лексических значений, а также некоторые известные лексико-семантические ресурсы.

# 1.Лексические отношения

Рассмотрим типы лексических отношений, которые являются наиболее востребованными в задачах автоматической обработки текстов.

**Синонимами** считаются слова, имеющие «общее ядро значения», но различающиеся «оттенками значения», поскольу известно, что абсолютных синонимов в естественных языках очень мало. Близкие по смыслу слова могут соответствовать друг другу по их денотатам, т.е. совпадает множество объектов действительности, которые могут именоваться этими словами. При этом слова могут отличаться, например, коннотациями,

т.е. положительными или отрицательными ассоциациями, связанными с этими словами. Или слова могут различаться по стилю речи, в которых это слово употребляется (книжный, научный, художественный, разговорный и др.)

Различия между синонимами могут быть:

- стилистические (*жена — супруга);*
- эмоционально-экспрессивные (*лошадь — кляча)*
- профессиональные *(воспаление легких — пневмония)*

Обычно синонимами считаются слова одной и той же части речи. Это связано с проверкой синонимии посредством взамозаменяемости в предложениях. Однако можно рассмотреть и **синтаксические синонимы (или дериваты**), которые различаются только принадлежностью к определенной лексико-грамматической категории, например, глагол доверять и существительное доверие.

**Антонимы** – это слова, которые имеют противоположные или обратные, но не противоречащие значения. Основными типами антонимов являются следующие:

- Начинать – переставать (*взвалить – свалить*)
- действие – уничтожение результата действия (*приклеить – отклеить*)
- Р – не Р (*влажный – сухой*)
- Больше – меньше (*узкий – широкий*)

Важным типом отношений между словами является отношение **гипоним-гипероним**, где гипероним – это слово с более широким значением. Для определения отношения можно применить следующие диагностические проверки:

Х рассматривается как гипоним У-а , если выполняются два условия::

1) из утверждения «А – это Х», следует, утверждение «А – это У»,

2) из утверждения «А – это У» не следует утверждение «А – это Х»

Например,

«Это – собака», значит «Это-животное». - «Это – животное» не следует «это собака»

«Это – жеребец», значит, «Это – лошадь». «Это – лошадь» не следует «это - жеребец».

«Оно – алый» , значит «Он – красный ». «Он красный» не следует «он алый»

Отношение часть-целое (меронимия) представляет собой скорее совокупность несколько отличающихся отношений, чем четкое отделяемое отношение. В качестве определения меронимии, которое однако исключает некоторые очевидные случаи отношения часть-целое, может служить следующее положение:

X является меронимом Y тогда и только тогда, если предложения вида *Y имеет X (или Хы) и X – это часть Y* являются нормальными для X и Y, интерпретируемых как родовые понятия.

Наиболее центральным типом этого отношения являются физические объекты. Сущности, длящиеся во времени, могут иметь части. Мы можем ссылаться на них как на стадии, фазы, этапы. Сущности, такие как группы, классы и коллекции, состоят в отношении меронимии со своими элементами, например, это такие слова, как *племя, команда, комитет, семья, оркестр, суд, отряд и др*.

Лексические отношения позволяют сделать правильный вывод по тексту в следующих примерах:

- Домашние животные запрещены => собаки запрещены (гипоним)
- Ресторан в Японии=> ресторан в Азии (целое или холоним)
- Ресторан в Японии ≠ Ресторан в Китае (ко-гипонимы, т.е. гипонимы одного и того же гиперонима)
- Хороший ресторан ≠ плохой ресторан (антонимы)

При ответе на вопросы по текстам может пригодиться знание отношений часть-целое и отношений ко-гипонимов, как например, в следующем примере.

- Когда Дональд Трамп посещал Францию?
- Правильный ответ: Трамп посещал Париж в сентябре (географическая часть)
- Информация, которую даже можно и не рассматривать в качестве правильного ответа: Трамп посещал Испанию в октябре (ко-гипоним)

В настоящее время получили большое распространение векторные представления слов (word embeddings), вычисляемых на больших объемах текстовых данных. Такие векторные представления дают возможность вычисления семантической близости слов, хорошо коррелирующие с человеческими представлениями о семантической близости.

Однако в настоящее время такие представления не могут с достаточной уверенностью различать разные типы отношений между словами. Например, векторные представления слов дают очень высокое сходство между антонимами, но в системах анализа тональности, вопросно-ответных системах является очень различение антонимов для корректной выдачи. Кроме того, векторные представления слов могут иногда выдавать высокую семантическую близость для не похожих по смыслу слов, которую трудно интерпретировать.

# 2. Тезаурусы типа WordNet

Лингвистический ресурс WordNet для английского языка разработан в Принстонском университете США и представляет лексику английского языка в виде семантической сети, которая содержит значения слов и лексические отношения между ними. WordNet свободно доступен в Интернет, и на его основе были выполнены тысячи экспериментов в области информационного поиска и автоматической обработки текстов. WordNet версии 3.0 охватывает приблизительно 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч понятий, или совокупностей синонимов (synset); общее число пар лексема-значение насчитывает 200 тысяч. В разных странах предприняты усилия по созданию ресурсов для своих языков по модели WordNet (ворднеты).

Основным отношением в WordNet является отношение синонимии. Наборы синонимов – синсеты – основные структурные элементы WordNet. Понятие синонимии базируется на критерии, что два выражения являются синонимичными, если замена одного из них на другое в предложении не меняет значения истинности этого высказывания.

Поскольку требование заменяемости двух близких по смыслу слов во всех возможных контекстах является слишком жестким, поэтому на практике используется значительно более слабое утверждение, что синонимы WordNet должны быть взаимозаменимы в значительном множестве контекстов. Например, замена *plank* (доска, планка) для слова *board (доска)* редко меняет значение истинности в контексте плотницкого дела, но существуют контексты, где такая замена не может считаться приемлемой.

Большинство синсетов снабжены толкованием, подобным толкованиям в традиционных словарях, — это толкование рассматривается как одно для всех синонимов синсета. Если слово имеет несколько значений, то оно входит в несколько различных синсетов. Синсет, рассматривается авторами, как представление лексикализованного понятия (концепта) английского языка.

Тезаурус WordNet включает слова четырех частей речи (существительные, прилагательные, глаголы и наречия) и разделен на четыре семантические сети в соответствии с этими частями речи. Синсеты каждой части речи в WordNet имеют свой набор отношений. Предполагается, что разделение синсетов по разным частям речи соответствует психолингвистическим экспериментам, которые показывают, что представление

информации о прилагательных, существительных, глаголах и наречиях устроено в человеческой памяти по-разному.

Отметим, что последние версии WordNet включают также отношения словообразования (derivationally related form), которые связывают между собой лексические единицы разных частей речи. Например, глагольный синсет {*change, alter, modify*} (*изменять, модифицировать*) связан такими отношениями с существительными *changer, change, alteration, modification*, и прилагательными *alterable, modifiable*.

В различных компьютерных приложениях чаще всего используется семантическая сеть существительных, между которыми установлены отношения синонимии, антонимии, гипонимии (гиперонимии), меронимии (часть-целое), поэтому рассмотрим представление существительных более подробно.

Основным отношением между синсетами существительных является родовидовое отношение, при этом видовой синсет называется гипонимом, а родовой — гиперонимом. Это отношение рассматривается как транзитивное, т.е. если В является гиперонимом для А, а С – это гипероним для В, то предполагается, что С является гиперонимом для А. Так, например, если указано, что ворона – это птица, а птица – это животное, то можно сделать вывод, что ворона – это животное. Синсет X называется гипонимом синсета Y, если носители английского языка считают нормальными предложения типа «An X is a (kind of) Y”.

Таким образом, отношения между синсетами образуют иерархическую структуру. При построении иерархических систем на базе родовидовых отношений обычно предполагается, что свойства вышестоящих понятий наследуются на нижестоящие – так называемое свойство наследования. Таким образом, существительные в WordNet организованы в виде иерархической системы с наследованием; были сделаны систематические усилия, чтобы для каждого синсета найти его родовое понятие, его гипероним.

Среди отношений часть-целое, которые описаны в WordNet для существительных, дополнительно выделяются отношения быть элементом (member_of, например, дерево – лес), а также отношение быть материалом (substance_of: стекло – стеклянная посуда).

Считается, что меронимы могут наследоваться гипонимами, например, если крыло и клюв описаны как части птицы, то все виды птиц наследуют эти части.

Авторы подчеркивают, что одной из проблем описания отношений меронимии является то, что части описываются несколько выше, чем это необходимо. Например, часто утверждается, что колесо – это часть транспортного средства, но тогда сани не являются транспортным средством. Однако часто такая ситуация является следствием

того, что понятие необходимого уровня не лексикализовано в языке. Для данного конкретного примера WordNet вводит специальное дополнительное понятие {wheeled vehicle} – колесное транспортное средство.

Еще одним отношением, установленным для существительных, является отношение антонимии. Отношение антонимии является отношением между конкретными словами, не между синсетами. Кроме того, отношение антонимии не наследуется на синсеты-гипонимы. Предполагается, что отношение антонимии должно быть явным образом описано. Примерами отношений антонимии в WordNet являются следующие: *победа – поражение, счастье – несчастье, мужчина – женщина.*

Серьезное обсуждение возникло по поводу описания значений многозначных слов в WordNet. Было подсчитано показано, что среднее количество значений в WordNet больше, чем в традиционных лексикографических словарях. Во многих работах признается, что различия значений в WordNet слишком тонки для таких компьютерных приложений как машинный перевод, информационный поиск, классификация текстов, вопросно-ответные системы и др.

Особенно большое количество значений имеют глаголы и прилагательные. Так, глагол *give* имеет 44 значения, а прилагательное *good* – 21 значение. Часть выделенных значений сочетается только с узким набором слов, например, значение give19: Give19: give - (***give*** (as medicine); ``I gave him the drug'' ) – дать (как лекарство)

Эти проблемы привели к постановке вопроса о том, каким образом и какие типы значений многозначного слова могут быть объединены («кластеризованы») для целей работы в приложениях автоматической обработки текстов, когда для значений многозначного слова из кластера можно не делать различий, и это не приведет к снижению качества работы этого приложения. В результате проведенных исследований выяснилось, что кластеризация значений может проводиться на основе различных взаимоисключающих критериев (семантических, синтаксических, предметно-ориентированных), что говорит также и о разной значимости разных подразделений значений для конкретных приложений автоматической обработки текстов

# 3. Вычисление семантической близости текстов на основе тезауруса

Одной из задач автоматической обработки текстов является определение семантической близости между словами, которая обычно оценивается как величина от 0 до 1, при этом 1 – это максимальная семантическая близость.

Базовое предположение для вычисления семантической близости на основе тезауруса состоит в том, что чем короче путь в сети между синсетами, в которые входят слова, тем они ближе по смыслу. При этом обычно рассматриваются не произвольные пути, а пути по иерархии отношений гипоним-гипероним. Важным здесь является термин Least Common Subsumer, т.е. это наиболее ближайший гипероним, который встречается, если подниматься по гиперонимическим отношениям от двух слов, для которых нужно вычислить семантическую близость.

Отметим, что в данных подходах обычно предполагается, что путь длины 1 – это путь между синонимами одного и того же синсета. Одно отношение между синсетами будет рассматриваться уже как путь длины два.

Можно выделить следующие методы вычисления семантической близости между словами на основе путей тезауруса:

- - методы, использующие только длину пути между узлами в тезаурусе (PATH),
- - методы, использующие длину пути между узлами в тезаурусе и глубину узлов в иерархии,
- - методы, на основе меры «информационного содержания» (information content).

Меры на основе длин путей используют для вычисления близости только длину пути межу узлами сети. Одним из вариантов такой меры может быть следующая:

$$path\,(a,b) = \frac{1}{shortest\_hypernym\_path\,(a,b)}$$

Если учитывать только длину пути, то мы лишаемся информации о специфичности понятий. Понятия, лежащие на более глубоких уровнях иерархии, являются более специфичными и кажутся более семантически близкими друг к другу, чем более общие понятия. Поэтому были предложены меры, использующие дополнительно глубину иерархии, как, например, мера Wu-Palmer:

$$wup(a,b) = \frac{2 \cdot depth(LCS(a,b))}{depth(a) + depth(b)}$$

$$depth(x) = shortest\_hypernym\_path(x, root)$$

Мера на основе информационного содержания определяется как величина вероятности встретить пример понятия С в большом корпусе *P(C)*. Эта вероятностная функция обладает следующим свойством: если $C_1$ вид для $C_2$, то $P(C_1) \leq P(C_2)$. Значение вероятности для наиболее верхней вершины иерархии равно 1. Следуя обычной аргументации теории информации, информационное содержание понятия *C* может быть представлено как отрицательный логарифм этой вероятности:

$$IC(C) = -log(P(C)).$$

Чем более абстрактным является понятие, тем меньше величина его информационного содержания. Для решения задачи разрешения лексической многозначности, вводится понятие наименьшего общего вышестоящего (LCS = Least Common Subsumer). Вычисления близости между узлами может вычисляться по форме Lin

$$sim_{lin} = \frac{2 \cdot IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2)}$$

или по формуле Jcn

$$sim_{jcn} = \frac{1}{IC(C_1) + IC(C_2) - 2 \cdot IC(LCS(C_1, C_2))}$$

Для вычисления близости между словами могут использоваться также не только конкретные пути между синсетами слов, но и структура всего графа отношений ресурса. Одним из методов, который может быть использован таким образом, - это метод PageRank, изначально предложенный для определения авторитетности интернет-страниц в глобальных поисковых системах.

В алгоритме PageRank рассматривается ситуация случайного блуждания по сети ссылок (random walking). Пусть пользователь бродит по страницам случайным образом: начинает на случайно странице, на каждом шаге переходит на другую страницу по исходящей ссылке с равной вероятностью для всех исходящих ссылок. Тогда в пределе каждая страница получит рейтинг посещений, который может использоваться как показатель авторитетности страницы.

Однако на пути случайного блуждания могут встретиться тупиковые страницы, в которых случайное блуждание остановится. Поэтому предложен механизм телепортации, который состоит в том, что

- для тупиковой страницы с равной вероятностью делается переход на любую другую страницу в сети,

- с нетупиковой страницы, т.е. страницы, в которой есть исходящие ссылки, с некоторой заданной вероятностью (коэффициент телепортации, например 0.1 ) производится переход на случайную страницу сети, а с оставшейся вероятностью равновероятно делается переход по одной из исходящих ссылок.

PageRank узлов сети может быть вычислен с помощью итеративного алгоритма, используя следующую формулу:

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

eRank страницы, L() – количество исходящих ссылок страницы.

В начале каждому узлу сети присваивается случайное значение PageRank. После этого, начиная с любой страницы, по формуле пересчитывается PageRank этой страницы, используя PageRank'и остальных страниц.

Таким образом, чем чаще пользователь при описанном процессе случайного блуждания попадает на некоторую страницу, тем больше ссылок ведет к этой странице, и, следовательно, она более авторитетна.

Для определения семантической близости слов исходная мера PageRank может быть модифицирована следующим образом. Граф отношения между синсетами рассматривается как сеть для случайного блуждания. Отношения между синсетами рассматриваются как ссылки для случайного перехода. Для того чтобы вычислить наиболее близкие слова для заданного слова, необходимо выполнять телепортацию только на синсеты, в которых в качестве синонима есть заданное слово. Тогда все остальные синсеты, и, следовательно, слова, упорядочатся по мере мере близости (авторитетности) к заданному слову.

# 4. Подходы к созданию тезаурусов типа WordNet для других языков

Создание больших лексических ресурсов типа WordNet с нуля представляет собой сложную задачу, которая требует усилий в течение многих лет. При создании тезаурусов

для автоматической обработки текстов разработчики должны последовательным образом решать совокупность сложных лексикографических проблем, усугубляемых тем, что система отношений представлена формализованным образом, у основных отношений гипоним-гипероним предполагается транзитивность, а сам ресурс предполагается использовать в автоматическом режиме. Решаемые проблемы включают в себя:

- разбиение множества близких по смыслу слов на совокупности синсетов,
- способы представление словосочетаний,
- выделение систем значений многозначных слов.

Чтобы ускорить разработку ворднета для своего языка, первая версия такого ресурса может быть создана посредством автоматического перевода Принстонского WordNet на целевой язык, 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.

В настоящее время отмечается, что ворднеты для разных языков, сохраняя базовую структуру построения, могут значительно отличаться друг от друга по принципам включения слов и выражений в синсеты, набору семантических отношений между синсетами, интерпретации конкретных семантических отношений. Также в ворднетах может значительно различаться подходы к описанию полисемии, что приводит к более дробной или более крупной системе представления значений многозначных слов. Могут различаться и подходы к включению в ворднет словосочетаний.

В настоящее время развивается проект Open Multilingual Wordnet, целью которого является связать между собой существующие ворднеты, созданные для разных языков с открытой лицензией (http://compling.hss.ntu.edu.sg/omw/). В 2020 году в проекте доступны 35 ворднетов для разных языков, для синсетов каждого из которых установлены соответствия с синсетами Принстонского WordNet. Все эти ворднеты могут быть получены в виде единого файла в нескольких форматах. Также ко всем ворднетам проекта имеется доступ через пакет NLTK (http://www.nltk.org/howto/wordnet.html). Для подключения в проект ворднета нового языка необходимо связать синсеты данного языка с синсетами WordNet и представить данные в требуемом формате.

# 5. Тезаурусы типа WordNet для русского языка

Известно, по крайней мере, пять проектов создания тезауруса типа WordNet для русского языка. В проекте RussNet авторы создавали русский ворднет с нуля, руководствуясь принципами WordNet. В двух разных проектах были предприняты попытки произвести автоматический перевод WordNet на русский язык с сохранением всей исходной структуры тезауруса. Результаты одной из работ опубликованы (wordnet.ru) - и анализ порожденного таким образом тезауруса показывает, что требуется его значительное редактирование или использование более качественных алгоритмов.

Еще один проекте YARN (Yet Another Russian wordNet) был инициирован в 2012 и создавался на основе краудсорсинга, т.е. участия в работе по наполнению тезауруса большого количества участников (https://russianword.net/). В настоящее время YARN содержит значительное количество синсетов с незначительным количеством отношений между ними. Работа над проектом прекращена в 2018 году

Последний по времени тезаурус типа WordNet, RuWordNet, был создан автоматизированным преобразованием в соответствующую структуру другого тезауруса русского языка РуТез (https://www.labinform.ru/pub/ruthes/index.htm).

Тезаурус РуТез начал создаваться более 20 лет назад в качестве лексико-терминологического ресурса для приложений автоматической обработки текстов и информационного поиска текстов на русском языке. Структура РуТез представляет собой семантическую сеть понятий и отношений между ними. В отличие от WordNet РуТез не разделен на отдельные подсети по частям речи, а слова и выражения разных частей речи могут быть текстовыми входами одного и того же понятия (например, *красный, краснота, красный цвет*). Тезаурус РуТез имеет несколько другой набор отношений, чем WordNet, что связано с изначальным развитием модели тезауруса для работы в предметных областяъ, где менее существенны лексические отношения, а важны взаимоотношения между терминами.

В течение своего развития тезаурус использовался в различных приложениях и проектах с государственными организациями и компаниями. Принципы разработки тезауруса для автоматической обработки текстов по модели РуТез были неоднократно использованы для создания тезаурусов и онтологий в разнообразных предметных областях, включая Онтологию по наукам и технологиям ОЕНТ, онтологию в области авиации АВИА-ОНТОЛОГИЯ, Тезаурус по компьютерной безопасности, Банковский тезаурус (сделан по заказу Центрального Банка Российской Федерации) и др. Разработана единая методология создания тезауруса

(лингвистической онтологии) новой предметной области, на основе представительной коллекции документов этой области, включая автоматизированное извлечение терминологии, принципы ввода новых понятий-терминов в тезаурус и описания отношений. С 2013 года версия тезауруса РуТез (РуТез-lite) опубликована и доступна для некоммерческого применения.

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. Поэтому было принято решение об автоматизированном преобразовании тезауруса РуТез в структуру вида WordNet/

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

1) разделить сеть понятий тезауруса РуТез 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). Между разделенными синсетами, относящимися к одному исходному понятию тезауруса РуТез, были установлены отношения частеречной синонимии.

**Таблица 1**. Количественные характеристики тезауруса RuWordNet

| 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 РуТез и скорректированы в соответствием 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.

Синсеты прилагательных часто имеют отношения частеречной синонимии к синсетам существительных и глаголов. Например, синсет слова *строительный* имеет два таких отношения: к синсету существительных {*стройка, постройка, возведение, сооружение..*} и к синсету глаголов {*строить, построить, возводить ...*}.

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*},

- синсет {*легкий, легкий для выполнения, легкий для осуществления, нетрудный*} представлен как антонимичный к синсету {*тяжкий, трудный*, *тяжелый, трудный для выполнения, нелегкий ...* },

# 6. Википедия как многоязычный онтологический ресурс

Википедия известна как Интернет-энциклопедия, создаваемая пользователями на многих языках. Википедия может быть характеризоваться как продукт краудосорсинга, т.е. работы большого количества людей по всему миру. При этом Википедия позиционируется как «свободная энциклопедия», которая содержит так называемый свободный «контент»,

для которого не накладывается ограничений на использование и получение выгоды, распространение копий, улучшение исходного материала и распространение его производных продуктов[1].

Таким образом, применение знаний, представленных в Википедии, в различных задачах, включая приложения автоматической обработки текстов, не требует никаких дополнительных финансовых вложений от разработчиков. Поэтому с самого начала появления Википедии появились работы, которые используют ее контент, для решения разнообразных задач автоматической обработки текстов, включая автоматическую классификацию текстов, вычисление семантической близости текстов, распознавание смысловой близости поисковых запросов, извлечение ключевых слов из документов и др. В подходах на основе Википедии может использоваться как система связей между страницами Википедии, так и представление понятия как текстового вектора на основе соответствующей словарной статьи Википедии.

Также Википедия может быть рассмотрена как многоязычный компьютерный онтологический ресурс, который можно использовать для автоматической обработки текстов.

Каждая статья Википедии (страница) представляет информацию об определенном понятии или именованной сущности. Между понятиями, представленными в Википедии, имеются несколько типов отношений:

- отношения перенаправления связывают между собой конкретные языковые единицы и страницы Википедии, с каждым понятием может быть связано несколько слов и выражений, и таким образом слова и выражения, связанные с одним и тем же понятием, образуют рад синонимов, синсет, подобный синсету тезаурусов типа WordNet,
- отношения-значения (disambiguation pages) описывают многозначные языковые единицы, которые ведут к нескольким страницам Википедии, это также похоже на описание многозначности в ворднетах, в которых многозначные слова входят в состав разных синсетов,
- внутренние отношения между понятиями представлены как гиперссылки в тексте статьи, и ведут на близкие по тематике страницы, обеспечивая таким образом систему отношений между понятиями, представленными в Википедии,

[1] https://ru.wikipedia.org/wiki/%D0%A1%D0%B2%D0%BE%D0%B1%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D0%BA%D0%BE%D0%BD%D1%82%D0%B5%D0%BD%D1%82

- межязыковые отношения обеспечиваются тем, что имеются отношения между страницами Википедии, посвященными одному и тому же понятию, написанные на разных языках.
- категории – страницы Википедии могут связаны к одной или нескольким категориям, которые могут быть двух видов: категории-множества, например, C: Cities, которая содержит статьи про конкретные города (Москва, Нью-Йорк и т.д.), и категории-топики, например, C-City, которая содержит статьи по городской тематике (городское планирование, урбанизация и др.).

Таким образом, Википедия может рассматриваться как формализованное представление знаний о мире, т.е. ресурс онтологического типа. Она может быть представлена как граф, в котором вершинами являются Вики-страницы, а ребрами - отношения между Вики-страницами. В свою очередь, и ресурсы типа WordNet могут быть представлены в виде графа, в котором вершинами являются синсеты, а ребрами являются описанные семантические отношения.

Такие представления WordNet и Википедии позволяют ставить задачи объединения WordNet и Википедии в единый онтологический ресурс, в котором присутствует более широкое покрытие существующих понятий и экземпляров, чем в WordNet. В то же время WordNet вносит в структуру Википедии более строго описанные семантические отношения.

Склеивание лингвистических ресурсов, таких как тезаурусы типа WordNet и полуструктурированный ресурс типа Википедия дает возможность также для создания многоязычного ресурса в виде семантической сети. Одним из известных ресурсов такого рода в настоящее время является ресурс BabelNet.

Для сопоставления WordNet и английской Википедии в BabelNet используется следующая информация исходных ресурсов:

- все имеющиеся синсеты и значения слов текущей версии WordNet 3.0 вместе с их лексическими и семантическими отношениями,

- все имеющиеся содержательные статьи Википедии, которые рассматриваются как понятия Википедии, а также ассоциативные отношения между ними, извлеченные на основе гиперссылок между соответствующими страницами.

Для создания единого многоязычного ресурса необходимо:

- склеить соответствующие друг другу синсеты WordNet и понятия Википедии в так называемые синсеты BabelNet,

- нарастить многоязычные текстовые входы синсетов BabelNet за счет: a) установленных в Википедии ссылок между страницами одного и того же понятия на разных языках, b) использование системы автоматического перевода

-установить отношения между синсетами BabelNet, используя все отношения из WordNet, а также все ассоциативные отношения соответствующих понятий Википедии, при этом извлекаются отношения из вики-страниц на всех языках, с которыми идет работа в текущей версии BabelNet.

В настоящее время в ресурс BabelNet объединены и другие существующие многоязычные ресурсы, например словарь Wiktionary и ворднеты из проекта Open Multilingual Wordnet, что позволило включить в BabelNet слова и выражения из более, чем 270 языков. Ресурс позволил улучшить качество разрешения неоднозначности для разных языков, а также в близкой задаче entity linking, т.е. связывания сущностей, упомянутых в тексте, с каким-либо базовым ресурсом, например, Википедией.

# Заключение

В целом, нужно отметить, что лексико-семантические ресурсы (тезаурусы, лингвистические онтологии, семантические фреймы) являются востребованными ресурсами в области автоматической обработки текстов, с помощью процедур интеграции ресурсов увеличивается покрытие и мультиязычность такого рода ресурсов.

Лексико-семантичкксие ресурсы используются в таких приложениях, как:

- семантический анализ текстов,

- семантическое концептуальное индексирование в информационно-поисковых и информационно-аналитических системах,

- образование дополнительных специализированных ресурсов, например, ImageNet, SentiWordNet,

- создание обогащенных представлений текстов, интегрирующих знания в статистические методы (вероятностные тематические модели, распределенные представления слов. Так, современные представления слов в виде векторов имеют несколько проблем, включая смешение разных значений слов, сложности с представлением словосочетаний, а также есть проблема, что очень близкие синонимы, разные имена одного и того же объекта могут иметь достаточно разные представления

- в качестве источника признаков для систем машинного обучения, включая подходы в рамках глубокого машинного обучения и др.

# Литература к лекции


1. Азарова И. В., Синопальникова А. А., Яворская М. В. Принципы построения wordnet-тезауруса RussNet //Кобозева ИМ, Нариньяни АС, Селегей ВП (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог. – 2004. – С. 542-547.
2. Апресян Ю. Д. Лексическая семантика (синонимические средства языка). Избранные труды. Т. I. М.: Языки русской культуры, 1995.
3. Кобозева И. М. Лингвистическая семантика. Изд.4, УРСС, 2009.
4. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Москва 2010.
5. Agirre E., Soroa A. Personalizing pagerank for word sense disambiguation //Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). – 2009. – С. 33-41.
6. Bond F., Foster R. Linking and extending an open multilingual wordnet //Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2013. – С. 1352-1362.
7. Cruse D. A. et al. Lexical semantics. – Cambridge university press, 1986.
8. Loukachevitch N., Lashevich G., Dobrov B. Comparing Two Thesaurus Representations for Russian //Proceedings of Global WordNet Conference GWC. – 2018. – С. 35-44.
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).
10. Miller G. A. WordNet: An electronic lexical database. – MIT press, 1998.
11. Navigli R., Ponzetto S. P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network //Artificial Intelligence. – 2012. – Т. 193. – С. 217-250.
12. Vossen P. A multilingual database with lexical semantic networks // Dordrecht: Kluwer Academic Publishers. doi. – 1998. – Т. 10. – С. 978-94.