При обработке естественного языка, существует два подхода разрешения многозначности: поверхностный и глубокий.

В основе глубоких подходов лежит использование знаний о мире. Например, знание того, что « неодушевлённая, материальная вещь может быть зелёным в смысле цвета, но не может быть зелёным в смысле неопытности», позволяет определить, в каком смысле слово «зеленый» употреблено в данном контексте. Такие подходы не настолько результативны на практике, поскольку такой класс знаний о мире, пусть и возможно хранить в удобном для восприятия компьютера формате, покрывает очень небольшие области нашей с вами жизни и не совсем применимы ко всем исследованиям. Надо сказать, что и этот подход тоже не всегда работает, например, в предложении «Директор был такой зелёный» пользуясь знаниями, невозможно определить, в данном случае директор зелёный потому что он позеленел или потому что он неопытен — зачастую это возможно определить только исходя не из контекста, а логики и смысла всего текста.

Поверхностные же подходы не пытаются понять текст, они лишь опираются на анализ близлежащих слов, например: если рядом со словом «bass» присутствуют слова «sea» или «fishing», скорее всего, что в данном случае имеет место значение в биологическом смысле. Эти правила могут быть автоматически извлечены, используя корпус текстов с размеченными значениями слов. Этот подход, пусть и не покрывает по мощности предыдущий, по эффективности на практике легко его обгоняет. Однако, всегда существуют подводные камни, например, как в предложении «The dogs bark at the tree», в котором рядом со словом «bark» содержатся слова и «tree», и «dogs».

Выделены четыре основных метода разрешения многозначности:

* методы использующие тезаурусы, лексикографические базы данных – методы, основанные на знаниях;
* методы использующие размеченные корпуса текстов для тренировки классификатора – метода обучения с учителем(испытуемая система обучается посредством примеров «стимул-реакция»);
* методы использующие вторичные знания, такие как определения терминов в толкованиях слов или выровненный двуязычный корпус – методы обучения с частичным привлечением учителя;
* методы использующие неразмеченные корпуса данных - методы обучения без учителя.

### Методы, основанные на знаниях

Метод Леска - метод, основанный на использовании знаний словаря. Он основывается на гипотезе, что слова, находящиеся рядом в тексте, связаны друг с другом и эту связь можно наблюдать у определений слов и их значений. Два (или более) слова могут оказаться близкими, если у обоих из них будет обнаружена пара значений с наибольшим пересечением слов в их определениях в словаре. К примеру, словосочетание «pine cone», в определениях обоих в одного из значений присутствует такие слова как «evergreen» и «tree». Также, как альтернативу предыдущему способу, можно использовать глобальную связь между этими словами, подсчитав семантическую близость каждой пары значений в WordNet.

Как альтернативу вышеизложенным методам можно использовать общую семантическую близость значений слов, основанную на WordNet’e. Использование так называемых моделей управления («selectional preferences» или «selectional restrictions») также бывает довольно полезным. К примеру, используя знание, что слово «bass» в значении рыбы часто встречается со словом «cook» или «eat», мы можем разрешить многозначность в таком предложении как «I am cooking bass». Однако, создать подобные знания о мире чрезвычайно трудозатратно и практически невозможно.

# **Метод Леска**

Идея метода заключалась в поиске значения слова в списке словарных определений с учетом контекста, где это слово использовано. Основным критерием для выбора значения послужило следующее правило: заложенный в этом определении смысл должен был частично совпадать со смыслом значений соседних слов в контексте.

### Принцип работы алгоритма

Наглядное представление работы алгоритма Леска

1. Отделение контекста для рассматриваемого слова — не более 10 ближайших по расположению слов.
2. Поиск всех определений в словаре (любом другом подобном источнике знаний) для рассматриваемого слова.
3. Поиск слов из контекста в каждом найденном определении. Если какое-либо слово из контекста присутствует в определении, тогда этому определению дается балл (повышается процент верности определения для данного контекста).
4. В качестве наиболее вероятного значения выбирается то, для которого такое пересечение оказалось больше.

### Методы обучения с учителем

Все методы обучения с учителем основаны на предположении, что контекст рассматриваемого нами слова предоставляет достаточно информации для того, чтобы вычислить то, в каком значении оно в данном случае применено (а значит знания, полученные из словарей и тезаурусов, отсекаются как лишние). Все модели обучения с учителем применялись к проблеме WSD, включая связанные с ними техники, такие как выбор переменных, оптимизация параметров и смешанные модели. Метод опорных векторов и метод обучения на примерах показали себя как одни из наиболее высокоэффективных методов на сегодняшний день, возможно, потому что они могут справиться с многопараметрическими свойствами слов и контекстов. Однако, вышеперечисленные методы имеют в качестве узкого места требование иметь огромное количество вручную размеченных текстов для обучения, что, как уже говорилось, трудоёмко и дорого. Снова встаёт проблема обладания подобными размеченными корпусами.

### Методы частичного обучения с учителем

Метод бутстрэппинга является распространённым способом итеративного обучения и оценки классификатора для увеличения его эффективности. Алгоритм начинается с небольшого количества начальных данных (англ. seed data) для каждого слова: либо небольшое количество вручную введённых примеров контекстов либо пара безошибочных правил определения значения слова (к примеру, слово «play» в контексте слова «bass» почти всегда обозначает, что слово подразумевается в музыкальном значении). Эти данные используются для тренировки классификатора, применяя любой из вышеприведённых методов обучения с учителем. Затем, классификатор применяется на множестве уже неразмеченных текстов для извлечения большой тренирующей выборки, в которую включены только «надёжные» контексты. Процесс итеративно повторяется: каждый следующий классификатор обучается на соответствующем ему бо́льшем множестве контекстов — и повторяется до тех пор, пока весь корпус не покрыт либо пока не достигнуто максимальное количество итераций.

Другой метод использует большие объёмы неразмеченных текстов для получения информации о совместной встречаемости слов, что может значительно дополнить наши данные. Также, правильно выравненный билингвальный корпус может использоваться для разрешения кросс-языковой многозначности, так как многозначное слово в одном языке всегда переводится на другой язык в зависимости от своего значения, в котором оно употреблено. Этот метод в каком-то смысле тоже можно считать методом частичного обучения.

Все вышеперечисленные техники могут позволить адаптировать методы обучения с учителем к другим областям.

### Методы обучения без учителя

Данный вид методов — одна из наиболее сложных WSD-задач. Основным предположением этого метода является утверждение: «схожие значения встречаются в схожих контекстах» и таким образом они могут быть извлечены из текста с помощью кластеризации, используя некоторую меру схожести контекстов[[29]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote29). Тогда, новые контексты могут быть причислены к одному из ближайших кластеров. Производительность метода безусловно ниже других методов, однако сравнение несколько проблематично из-за необходимости проецирования полученных кластеров на имеющиеся в словаре значения. Если же проецирование не требуется, то можно произвести оценки кластеризации (включая энтропию и чистоту). Учёные возлагают большую надежду на то, что методы обучения без учителя смогут помочь превозмочь недостатки получения знаний, так как они не требуют решения чрезмерно трудоёмких задач по синтаксической и семантической разметке всего корпуса.

## **О процессе разрешения многозначности**

Вообще, под неоднозначностью (или многозначностью) языкового выражения или речевого произведения (текста) понимают наличие у него одновременно нескольких различных смыслов[[1]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote1). Учёные разделяют несколько типов подобной многозначности: лексическую, синтаксическую и речевую, однако термин «WSD» включает в себя разрешение именно лексической (смысловой).

О чём речь, можно понять из следующего примера с неоднозначным словом «ключ»:

1. ключ как инструмент для открывания
2. ключ как источник воды

а также 3 контекста:

1. *Ключ подошёл, дверь открылась*
2. *Я напился из ключа*
3. *Жизнь бьёт ключом*

Для человека является очевидным, что в первом предложении слово «*ключ*» используется в первом значении, во втором предложении — соответственно во втором значении, а в третьем — возможны варианты. Разработка алгоритмов, имитирующих подобную способность человека может подчас становиться сложнейшей задачей.

Процесс разрешения требует нескольких вещей: [системы словарных знаний](http://www.wikiwand.com/ru/%D0%AD%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D1%81%D0%BB%D0%BE%D0%B2%D0%B0%D1%80%D1%8C) для определения множества значений слов и [корпус](http://www.wikiwand.com/ru/%D0%9A%D0%BE%D1%80%D0%BF%D1%83%D1%81%D0%BD%D0%B0%D1%8F_%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) текстов для разрешения (в некоторых случаях могут потребоваться [иные источники знаний](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#.D0.92.D0.BD.D0.B5.D1.88.D0.BD.D0.B8.D0.B5_.D0.B8.D1.81.D1.82.D0.BE.D1.87.D0.BD.D0.B8.D0.BA.D0.B8_.D0.B7.D0.BD.D0.B0.D0.BD.D0.B8.D0.B9)).

### Другие методы

Также существуют другие методы, основанные на совершенно отличающихся от вышеперечисленных принципах:

* Определение доминантности значения слова (Determining Word Sense Dominance)[[30]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote30)[[31]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote31)[[32]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote32)[[33]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote33).
* Разрешение, основанное на темах (доменах) корпуса (Domain-Driven Disambiguation)[[34]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote34)[[35]](http://www.wikiwand.com/ru/%D0%A0%D0%B0%D0%B7%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BB%D0%B5%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B9_%D0%BC%D0%BD%D0%BE%D0%B3%D0%BE%D0%B7%D0%BD%D0%B0%D1%87%D0%BD%D0%BE%D1%81%D1%82%D0%B8#citenote35)
* WSD, использующее кросс-языковые данные (Cross-Lingual Evidence)

## **Внешние источники знаний**

Знания являются одними из ключевых моментов разрешения многозначности: они предоставляют данные, на которые опирается сам процесс разрешения. Эти данные могут быть как корпусы текстов, так и словари, тезурусы, глоссарии, онтологии и т. д.:

1. Структурированные источники:
   * Тезаурусы
   * Электронные словари
   * Онтологии
2. Неструктурированные:
   * Текстовый корпус: аннотированный значениями слов (sense-annotated corpora) и неаннотированный (raw corpora)
   * Знания о совместной встречаемости слов (Collocation resources)
   * Другие ресурсы, такие как: списки встречаемости слов, стоплисты, доменные теги и т.д.

Polysemy.pdf

Среди основных методов разрешения лексической многозначности выделяют: методы, использующие внешние источники информации, и методы, базирующиеся на машинном обучении, работающие на размеченных корпусах текстов. Также применяются комбинации этих методов По другой классификации, методы разрешения лексической многозначности различают по типу используемых внешних источников информации:  
структурированные источники данных(машиночитаемые словари, тезаурусы, онтологии). Тезаурусы содержат информацию об отношениях между словами, такими, как: синонимия, антонимия и другие. Классическим примером тезауруса и машиночитаемого словаря для английского языка является WordNet, в котором слова организованы в виде синсетов (от англ. synonym set, группа синнимов), отношения указаны между синсетами; неструктурированные источники данных в виде корпусов текстов делятся на (а) неразмеченные корпуса (raw corpora)и (б) синтаксически и/или семантически размеченные корпуса.

Далее будут представлены примеры методов и алгоритмов разрешения лексической многозначности, разбитые на группы:

• нейронные сети – многообещающие методы с богатой историей;

• бустинг как метод улучшения точности алгоритма обучения;

• лексические цепочки – построение последовательности семантически связанных

слов;

• метод ансамбля байесовских классификаторов и сочетаемостные ограничения на основе байесовских сетей;

• контекстная кластеризация – кластеризация контекстных векторов, где разные кластеры соответствуют разным значениям слова;

• кластеризация слов – это кластеризация семантически близких слов, при этом кластер соответствует некоторому значению.

**WSD на основе нейронных сетей, построенных по данным машиночитаемых словарей**

В типичной NN на вход подается слово, значение которого требуется установить, т. е. целевое (target) слово, а также контекст (фраза), его содержащий. Узлы выхода соответствуют различным значениям слова. В процессе обучения, когда значение тренировочного целевого слова известно, веса связующих узлы соединений (связей) настраиваются таким образом, чтобы по окончании обучения выходной узел, соответствующий истинному значению целевого слова, имел наибольшую активность. Веса соединений могут быть положительными или отрицательными и настраиваются посредством рекуррентных алгоритмов (алгоритм обратного распространения ошибки, рекуррентный метод наименьших квадратов и т. д.). Сеть может содержать скрытые (hidden) слои, состоящие из узлов, соединенных как прямыми, так и обратными связями.

**Топология сети.**

Целевое слово представлено узлом, соединенным активирующими связями со смысловыми узлами, представляющими все возможные значения слова, имеющиеся в словарных статьях. Каждый смысловой узел, в свою очередь, соединен активирующими связями с узлами, представляющими слова в словарной статье, соответствующей толкованию данного значения. Процесс соединения повторяется многократно, создавая сверхбольшую сеть взаимосвязанных узлов. В идеале сеть может содержать весь словарь.

Алгоритм.

При запуске сети первыми активируются узлы входного слова (согласно принятой кодировке). Затем каждый входной узел посылает активирующий сигнал своим смысловым узлам, с которыми он соединен. В результате сигналы распространяются по всей сети в течение определенного числа циклов. В каждом цикле узлы слова и его значений получают обратные сигналы от узлов, содиненных с ними. Узлы конкурирующих значений посылают взаимно подавляющие сигналы. Взаимодействие сигналов обратной связи и подавления, в соответствии со стратегией .победитель получает все., позволяет увеличить активацию узлов-слов и соответствующих им правильных узлов-значений, одновременно уменьшая активацию узлов, соответствующих неправильным значениям. После нескольких десятков циклов сеть стабилизируется в состоянии, в котором активированы только узлы-значения с наиболее активированными связями с узлами-словами. При обучении сети используется метод обратного распространения (back propagation).

**Бустинг**

Бустинг – это общий и доказуемоэффективный метод получения очень точного правила предсказания путем комбинирования грубых и умеренно неточных эмпирических правил.

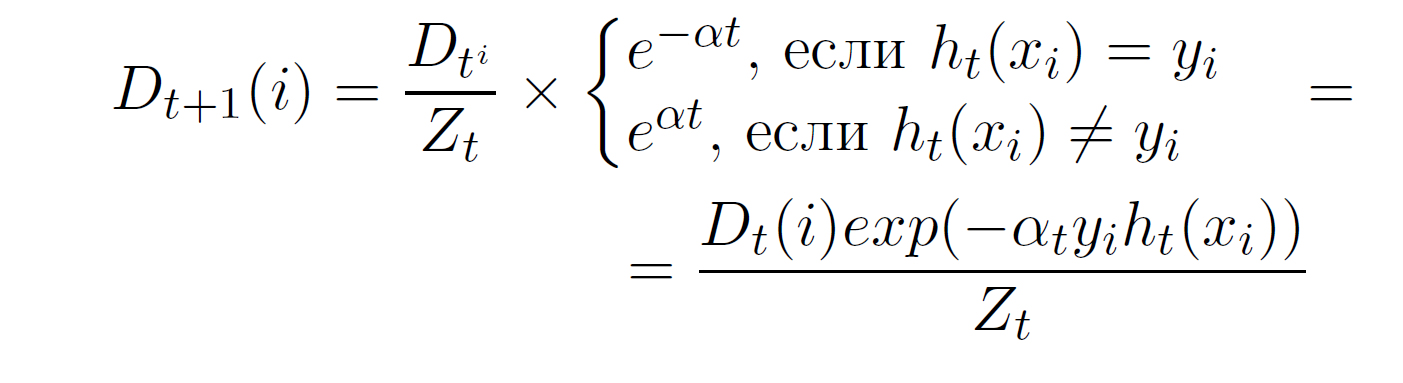
AdaBoost является адаптивным алгоритмом, поскольку он может адаптироваться к уровням ошибок отдельных слабых гипотез.

На вход алгоритма поступает обучающая выборка (𝑥𝑖; 𝑦𝑖); ..; (𝑥𝑚; 𝑦𝑚), где каждый элемент 𝑥𝑖 принадлежит некоторому домену или признаковому пространству 𝑋 и каждая метка 𝑦𝑖 принадлежит некоторому набору ме-

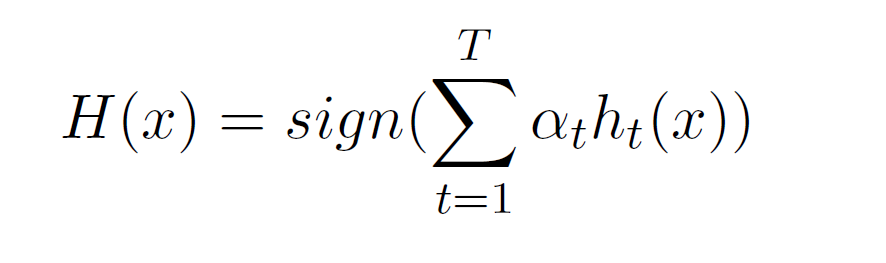
ток 𝑌 . Для каждого обучающего примера 𝑖 вес распределения для целых 𝑡 обозначается 𝐷𝑡 (𝑖), где 𝑡 – это шаг алгоритма. За начальное распределение весов принимается 𝐷1(𝑖) = 1/𝑚. Пусть метки принимают значения из множества 𝑌 = {−1, 1}. Далее на каждом шаге 𝑡, где 𝑡 = 1 . . . 𝑇,выполняется обучение с использованием текущего распределения 𝐷𝑡, после чего строится слабая гипотеза ℎ𝑡 : 𝑋 → {−1; 1} с ошибкой первого рода

𝜀𝑡 = Σ(𝑖 : ℎ𝑡(𝑥𝑖) ̸= 𝑦𝑖)𝐷𝑡(𝑖), по которой выбирается уровень значимости

𝛼𝑡 = (½)𝑙𝑛(1 – 𝜀𝑡/𝜀𝑡) и строится новое распределение для следующего шага



Конечная гипотеза 𝐻(𝑥) – это среднее из большинства решений 𝑇 слабых гипотез, где 𝛼𝑡 – вес, присвоенный гипотезе ℎ𝑡.



Идея алгоритма заключается в определении набора весов для обучающей выборки. Первоначально все веса примеров устанавливаются равными, но в каждом цикле веса неправильно классифицированных по гипотезе ℎ𝑡 примеров увеличиваются. Таким образом получаются веса, которые относятся к сложным примерам. Основное теоретическое свойство

AdaBoost – это способность алгоритма уменьшать ошибку обучения.

AdaBoost обладает определенными преимуществами. Его быстро и просто запрограммировать. Он не имеет никаких параметров для настройки, за исключением количества циклов. Он не требует никаких предварительных

знаний о слабом обучаемом и поэтому может быть скомбинирован с любым методом для нахождения слабых гипотез. Недостатки метода заключаются в следующем. Фактическая производительность бустинга на конкретной задаче явно зависит от данных и слабообучаемого алгоритма. Теоретически бустинг может выполниться плохо, если данных недостаточно, слабые гипотезы слишком сложные или, наоборот, слишкомслабые. Также бустинг особенно восприимчивк шуму.