## **О процессе разрешения многозначности**

Под неоднозначностью/многозначностью языкового выражения понимают наличие у него одновременно нескольких различных смыслов. Многозначность подразделяется на следующие типы: лексическую, синтаксическую и речевую, однако термин «WSD» включает в себя разрешение именно лексической (смысловой). Например, слово “ключ” может употребляться в одном из следующих значений: ключ как инструмент для открывания и ключ как источник воды.

Процесс разрешения требует нескольких вещей: системы словарных знаний для определения множества значений слов и корпус текстов для разрешения. 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.

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), неструктурированные источники данных в виде корпусов.

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

* методы, основанные на использовании тезаурусов, словарей и т.д.;
* методы, использующие нейронные сети;
* бустинг;
* 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.

**Методы использующие тезаурусы, лексикографические базы данных.**

В качестве примера одного из данных методов рассмотрим метод Леска, который основан на поиске значения слова в списке словарных определений с учетом контекста, в котором используется данное слово. Основным критерием при выборе значения является следующее правило: заложенный в этом определении смысл должен был частично совпадать со смыслом значений соседних слов в контексте.

### Метод леска можно разбить на следующие шаги:

1. Для исходного слова выделяется контекста, размер которого не более 10 ближайших по расположению слов;
2. Для исходного слова осуществляется поиск всех определений в словаре;
3. Сопоставление слов из контекста с каждым найденным определением. В случае если какое-либо из контекста слово присутствует в определении, то этому определению дается балл;
4. Наиболее вероятным значением является то, определение которого набрало наибольшее количество баллов.

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

В типичной нейронной сети на вход подается слово, значение которого требуется установить, т. е. целевое слово, а также контекст, его содержащий. Узлы выхода соответствуют различным значениям слова. В процессе обучения, когда значение тренировочного целевого слова известно, веса связующих узлы соединений 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 слои, состоящие из узлов, соединенных как прямыми, так и обратными связями.

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.

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, где каждый элемент  принадлежит некоторому домену или признаковому пространству 𝑋 и каждая метка  принадлежит некоторому набору меток 𝑌 . Для каждого обучающего примера 𝑖 вес распределения для целых 𝑡 обозначается , где 𝑡 – это шаг алгоритма. За начальное распределение весов принимается .Пусть метки принимают значения из множества 𝑌 = {−1, 1}. Далее на каждом шаге 𝑡, где 𝑡 = 1…𝑇, выполняется обучение с использованием текущего распределения , после чего строится слабая гипотеза  : 𝑋 → {−1; 1} с ошибкой первого рода , по которой выбирается уровень значимости  и строится новое распределение для следующего шага: 

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



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  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, слишком слабые. Также бустинг особенно восприимчив к шуму.

**Использование лексических цепочек для реферирования текстов**

Метод построения лексических цепочек включает шаги:

1. Выбирается набор слов-кандидатов на включение в цепочки (существительные и составные существительные).
2. По словарю строится список всех значений для каждого слова-кандидата.
3. Для каждого значения каждого слова-кандидата находится расстояние до каждого слова во всех уже построенных цепочках (слово в цепочке имеет строго определенное значение, задаваемое другими словами в той же цепочке). Между двумя словами есть отношение, если мало расстояние между этими словами в тексте или между значениями этих слов существует путь в тезаурусе WordNet. Выделяют три вида отношений:
4. Extra-strong отношение существует для слов, повторяющихся в тексте. Повтор может быть на любом расстоянии от первого употребления слова.
5. Strong отношение определено между словами, связанными отношением в WordNet. Два таких слова должны находиться в окне не более семи предложений.
6. Medium-strong отношение указывается для слов, синсеты которых находятся на расстоянии больше одного вWordNet (но есть еще и дополнительные ограничения на путь между синсетами). Слова в тексте должны находиться в пределах трех предложений.
7. Слово-кандидат добавляется в цепочки, со словами которых найдена связь. Смысловая неоднозначность устраняется, в цепочку добавляется не просто слово, а его конкретное значение (благодаря выбору значения в словаре на шаге 2).

Для выбора приоритетной цепочки (для вставки слова-кандидата) отношения упорядочены так: extra-strong, strong, medium-strong. Цепочки можно выбирать жадным алгоритмом, при этом слово-кандидат попадает ровно в одну цепочку и после этого выбор уже не может быть изменен, даже если последующий текст покажет ошибочность первоначального решения. Так же приоритетную цепочку можно выбирать по следующей схеме, требующей рассмотрения всех возможных цепочек. Таким образом, будут сформированы цепочки с учетом всех возможных значений слов с последующим выбором наилучшей цепочки.

**Разрешение лексической многозначности методом ансамбля байесовских классификаторов.**

Наивный байесовский классификатор – это простой вероятностный классификатор на основе применения теоремы Байеса. Для различения значений учитывается совместная встречаемость слов в окне заданного размера в текстах корпуса.

При разрешении лексической многозначности, представленном в виде задачи обучения с учителем, применяют статистические методы и методы машинного обучения к размеченному корпусу. В таких методах словам корпуса, для которых указано значение, соответствует набор языковых свойств.

Подход основан на объединении ряда простых классификаторов в ансамбль, который разрешает многозначность с помощью голосования простым большинством голосов. В проблеме разрешения лексической многозначности существует понятие контекста, в котором встречается многозначное слово. Этот контекст представляется в виде

функции переменных , а значение многозначного слова представлено в виде классификационной переменной S. Все переменные бинарные. Переменная, соответствующая слову из контекста, принимает значение “ИСТИНА”, если это слово находится на расстоянии определенного количества слов слева или справа от целевого слова. Совместная вероятность наблюдения определенной комбинации переменных контекста с конкретным значением слова выражается следующим образом: , где и – параметры данной модели. Для оценки параметров достаточно знать частоты событий, описываемых взаимозависимыми переменными. Эти значения соответствуют числу предложений, где слово, представляемое, встречается в некотором контексте многозначного слова, упомянутого в значении 𝑆. Если возникают нулевые значения параметров, то они сглаживаются путем присвоения им по умолчанию очень маленького значения. После оценки всех параметров модель считается обученной и может быть использована в качестве классификатора.

Контекст представлен в виде bagof-words (модель “мешка слов”). В этой модели выполняется следующая предобработка текста: удаляются знаки препинания, все слова переводятся в нижний регистр, все слова приводятся к их начальной форме (лемматизация). Контексты делятся на два окна: левое и правое. В первое попадают слова, встречающиеся слева от неоднозначного слова, и, соответственно, во второе – встречающиеся справа. Окна контекстов могут принимать 9 различных размеров: 0, 1, 2, 3, 4, 5, 10, 25 и 50 слов. Первым шагом в ансамблевом подходе является обучение отдельных наивных байесовских классификаторов для каждого из 81 возможных сочетаний левого и правого размеров окон. Наивный байесовский классификатор (𝑙, 𝑟) включает в себя слов слева от неоднозначного слова и 𝑟 слов справа. Исключением является классификатор (0, 0), который не включает в себя слов ни слева, ни справа. В случае нулевого контекста классификатору присваивается априорная вероятность многозначного слова (равная вероятности встретить наиболее употребимое значение). Следующий шаг при построении ансамбля – это выбор классификаторов, которыестанут членами ансамбля. 81 классификатор группируется в три общие категории, по размеру окна контекста. Используются три таких диапазона: узкий (окна шириной в 0, 1 и 2 слова), средний (3, 4, 5 слов), широкий (10, 25, 50 слов). Всего есть 9 возможных комбинаций, поскольку левое и правое окна отделены друг от друга. Например, наивный байесовский классификатор (1, 3) относится к диапазону категории (узкий, средний), поскольку он основан на окне из одного слова слева и окне из трех слов справа. Наиболее точный классификатор в каждой из 9 категорий диапазонов выбирается для включения в ансамбль. Затем каждый из 9 членов классификаторов голосует за наиболее вероятное значение слова с учетом контекста. После этого ансамбль разрешает многозначность путем присвоения целевому слову значения, получившего наибольшее число голосов.

**Контекстная кластеризация**

Каждому вхождению анализируемого слова в корпус соответствует контекстный вектор. Выполняется кластеризация векторов, где разные кластеры соответствуют разным значениям слова. Алгоритмы кластеризации полагаются на дистрибутивную гипотезу в соответствии с которой слова, употребляемые в схожих контекстах, считаются близкими по смыслу.

При решении задачи различения значений используются контекстные вектора: если целевое слово встречается в тестовых данных, то контекст этого слова представляется в виде вектора контекста. Вектор контекста — это средний вектор по векторам свойств каждого из слов контекста. Вектор свойств содержит информацию о совместной встречаемости данного слова с другими словами, этот вектор строится по данным корпуса текстов на этапе обучения.

Первоначально строится матрица совместной встречаемости слов по данным обучающего корпуса. Вектор свойств (строка матрицы) содержит информацию о совместной встречаемости данного слова с другими. После создания матрицы выполняется разделение тестовых данных, т. е. группировка примеров употреблений (фраз) с целевым словом. Каждому слову в примере употребления в тестовых данных соответствует вектор свойств из матрицы встречаемости. Средний вектор свойств по всем словам соответствует вектору контекста. Таким образом, набор тестовых данных, включающих употребление исследуемого слова, преобразуется в набор контекстных векторов, каждый из которых соответствует одному из употреблений целевого слова.

Различение значений происходит путем кластеризации контекстных векторов с помощью разделяющего (partitional) или иерархического “сверху вниз” алгоритма кластеризации. Получающиеся кластеры составлены из употреблений близких по значению фраз, и каждый кластер соответствует отдельному значению целевого слова. Векторы свойств, полученные по небольшому корпусу текстов, имеют очень малую размерность (несколько сотен), что не позволяет полностью описать закономерности совместной встречаемости слов. Для решения этой проблемы векторы свойств слов расширяются содержательными словами (content words), извлеченными из словарных толкований разных значений данного слова.

Данный метод может быть полезен при различении значений слов без учителя при небольшом количестве обучающих данных.