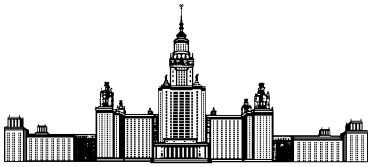
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Отчет о выполнении задания преддипломной практики

«Разработка и реализация компьютерной лаборатории оценки качества семантической близости слов»

Выполнил:

студент 4 курса 417 группы

Сенин Александр Николаевич

Научный руководитель:

к.ф.-м.н.

Майсурадзе Арчил Ивериевич

Москва, 2020

Содержание

Введение

Научный интерес автора во многом представляет задача многомерного шкалирования. Вкратце напомним ее. Известно, что существует некоторая коллекция из N объектов произвольной природы, на которой определена некоторая функция расстояния (distance function):

d_{ij} := расстояние между i-м и j-м объектом

Таким образом, про коллекцию объектов известна лишь матрица попарных различий (dissimilarity matrix): $D = \{d_{ij}\}_{i,j=1}^{N}$

Задача многомерного шкалирования: по матрице различий D найти такие N векторов $x_1, ..., x_N \in \mathbb{R}^p$, что $||x_i - x_j|| \approx d_{ij}$. В общем случае строго не оговаривается, что подразумевается под «приближением» расстояний: может рассматриваться в том числе и сохранение отношения порядка. Размерность полученного представления p обычно выбирается небольшой, например, $p = 2, 3$ для задач визуализации.

Так, например, основной вопрос курсовой работы автора заключался в решении задачи многомерного шкалирования с помощью нейронных сетей. Основной же вопрос выпускной квалификационной работы предварительно состоит в рассмотрении задачи многомерного шкалирования в случае неполной входной матрицы различий (причем ожидается, что пропусков будет много, возможно сильно больше, чем позиций с известными значениями).

Постановка задачи

Чтобы у задачи был контекст, была выбрана конкретная прикладная область — дистрибутивная семантика. Дистрибутивная семантика — это область лингвистики, которая занимается вычислением степени семантической близости между лингвистическими единицами на основании их распределения (отсюда дистрибутивность в названии) в больших массивах лингвистических данных (текстовых корпусах). Лингвистическими единицами в простейшем случае выступают слова. Математический подход к решению задачи нахождения семантической близости обычно заключается в представлении слов многомерными векторами.

Ожидается, что в результате решения некоторой задачи многомерного шкалирования будут получены некоторые представления слов. Интересует вопрос, как оценить полученные представления? Насколько качественные они получились с точки зрения дальнейшего использования?

Для оценки качества были определены четыре важных свойства представлений:

- Семантическая близость
- Сохранение аналогии
- Метричность
- Интерпретируемость

Остановимся подробнее на каждом свойстве. Неформально, свойство семантической близости означает близость в математическом смысле у полученных представлений для близких по смыслу слов. Отдельный вопрос, что понимать под близостью слов по смыслу. Чаще всего, выделяют два понятия: близость в смысле похожести (similarity) и связанность (relatedness). Объясним разницу этих понятий на примере. Нам знакомо понятие слов синонимов. По сути, это разные словесные описания одной и той же сущности. Далее, существуют слова, описывающие очень близкие вещи или категории, например, crocodile и alligator. По сути, такие слова можно считать почти синонимами, очень близкими в смысле похожести (similarity). Продолжая эту логику, слова car и crash не будут считаться близкими, так как описывают непохожие сущности. Однако, вполне естественно заметить, что слова автомобиль и авария в определенном смысле связаны (related). Это слова одной темы, не зря связанность часто определяют как близость тем или доменов (topical similarity, domain similarity). Еще пример, для слов car и train естественно говорить о близости, так как объекты, которые обозначают эти слова во-многом похожи (функция, материал, движение, колеса, окна и т.д.). Напротив, для слов Freud и psychology естественно говорить о связанности.

Основная идея свойства сохранении аналогий (см. Word Analogy Task) заключается в следующем: задаются некоторые два связанных слова a и b, и некоторое другое слово c. По полученным представлениям тем или иным образом строится новое слово

d, связанное с c также, как a связано с b. Таким образом, сохранение аналогий заключается в способности представлений строить слова по аналогии. Например, если задать в качестве слова $a =$ «man», а слова $b =$ «woman», то для слова $c =$ «uncle» естественно ожидать построения слова «aunt».

Чтобы тщательнее оценить качество построенных представлений, можно заодно ожидать функцию расстояния, которая предполагается для работы с этими представлениями. В случае, когда такой функции не предоставлено, можем считать расстояние на представлениях принятым в этом случае косинусным расстоянием. Если же функция расстояний предоставлена, то можем проверить ее метричность, то есть выполнение аксиом метрики (неотрицательность, симметричность, неравенство треугольника) хотя бы на заданных векторных представлениях.

Формализация того, что будем понимать под интерпретируемостью, пока обсуждается. Во-многом свойство сохранения аналогий покрывает некоторый запрос на интерпретируемость. Аналогично, что касается метричности. Для начала будем считать, что нам приходят только представления, без функции расстояния.

Теперь, когда стало ясно, как можно оценивать качество представлений, обсудим задание практики. Для оценивания семантической близости принято выделять заранее размеченные экспертами наборы данных, которые воспринимаются как эталон, будем их называть «золотыми стандартами». Основная часть таких наборов данных — набор пар слов (w_1, w_2), каждой паре соответствует оценка близости или связанности, в зависимости от метода построения набора данных. Не менее важная часть, хотя и не всегда встречающаяся — дополнительные метки слова в каждой паре (l_1, l_2). В нашем случае, метки будут нести смысл частей речи.

Первый этап задания — собрать и проанализировать наиболее популярные «золотые стандарты», выявить их особенности и недостатки, исправить возможные дефекты, изучить методы их построения. Далее требуется разработать программный интерфейс для взаимодействия с такими наборами данных. Для представления «золотых стандартов» на диске разработать стандартизированный формат, добавить в программу функционал считывания стандарта в память из стандартизированного формата и записи в этот же формат. Анализ данных в выбранном стандартизиро-

ванном виде ожидается с помощью соответствующего программного функционала.

Второй этап задания — подготовить собранные данные для дальнейшего использования. Во-многом сюда относится разработка стандарта для хранения данных на диске, о котором шла речь ранее. Помимо этого, требуется предложить наиболее удачный способ разделения «золотого стандарта» на обучающую выборку и контрольную. Далее, нужно программно реализовать выбранный способ разделения.

Третий этап задания — на основе имеющихся обработанных, приведенных к общему виду данных выбрать способ оценивания качества представлений, написать его программную реализацию. Весь код, написанный в ходе выполнения, нужно собрать в модуль, который будем называть «компьютерной лабораторией оценки качества семантической близости слов». Классы и методы в модуле сопроводить поясняющими комментариями.

Выбор и анализ «золотых стандартов»

Название	Язык	Количество пар	Диапазон оценки	Значения оценки
WordSim353	Английский	353	[0; 10]	Вещественные
SimLex-999	Английский	999	[0; 10]	Вещественные
The MEN Test Collection	Английский	3000	[0; 50]	Целые

Таблица 1: Общие характеристики «золотых стандартов».

Первоначально выбраны три самых популярных набора данных, связанных с оцениванием качества представлений. Общие характеристики см. в Таблице 1.

Подробнее остановимся на происхождении этих датасетов. Мы уже обсуждали принципиальную разницу близости (similarity) и похожести (relatedness). «Золотые стандарты» обычно получают путем усреднения оценок экспертов разметчиков. Перед началом процедуры разметки экспертам предоставляют инструкцию, в которой описан подход к оцениванию похожести или связанности слов.

- WordSim353 — самый старый из представленных наборов данных, выпущен Evgeniy Gabrilovich в 2002 году. 353 пары слов в датасете были разбиты на

две части: часть из 153 пар была размечена 14 экспертами, оставшаяся часть из 200 слов размечена 16 экспертами. Каждого эксперта просили оценить по шкале от 0 до 10 связанность (relatedness) слов в паре, а затем оценку усредняли между экспертами. В оригинальном датасете представлены оценки каждого конкретного эксперта, что можно использовать для получения лучшей оценки, чем обычное усреднение. Например, можно учитывать оценку каждого эксперта с весом, корректирующим смещение оценок эксперта в ту или иную сторону. Опционально, существует расширение датасета, с проставлением метки каждой паре об отношениях между словами в паре: identical tokens, synonym, antonym, hyponym, hyperonym, sibling terms, first is part of the second one, second is part of the first one, topically related.

Значительный недостаток этого набора — использование только существительных. Кроме этого, в инструкции не объясняется разница между similarity и relatedness, поэтому есть некоторая путаница между этими понятиями, следствием чего является нечувствительность к этой разнице в оценках датасета. Существуют модификации датасета с делением пар на similar и related.

- The MEN Test Collection — самый крупный из представленных набор данных с 3000 парами слов, собранный Elia Bruni в 2012 году. В этом наборе уже представлены слова трех частей речи: существительные, глаголы и прилагательные (причем внутри пары могут встречаться слова разных частей речи). Оценки парам получены другим способ, нежели в WordSim353: вместо того, чтобы просить эксперта оценить в некоторой шкале связанность в паре, эксперту предлагают выбирать между двумя парами, в какой слова более связаны. Оценка каждой паре получается в результате работы одного эксперта, здесь могут быть потенциальные проблемы, отсутствует согласование мнений нескольких экспертов. Каждому эксперту предоставляется пара слов и 50 случайно выбранных других пар, а затем от эксперта ожидается 50 сравнений, в какой паре слова более связаны. В результате, в случае наиболее свзяанных слов в паре, эксперт во всех случаях выберет эту пару в сравнении. Таким образом, пара получит наибольшую оценку 50.

Ключевой недостаток этого набора данных остается прежним — близость (similarity) считается авторами частным случаем связанности (relatedness). Экспертам вновь не объясняется разница между этими понятиями, поэтому в наборе могут возникать артифакты, связанные с нечувствительностью к близости и связанности.

- SimLex-999 — самый новый из представленных наборов данных, выпущен в 2014 году. Изначально, датасет задумывался с целью решить ключевую проблему WordSim и MEN — нечувствительность к разнице близости и связанности. Принцип получения оценок из мнений экспертов здесь такой же, как в WordSim353, но экспертам теперь явно в инструкции объясняется разница между similarity и relatedness, экспертов просят оценить именно similarity. Например, для similar слов 'coast' - 'shore' оценки 9.00 (SimLex-999) 9.10 (WordSim353), а для related слов 'clothes' - 'closet' оценки 1.96 (SimLex-999) 8.00 (WordSim353). В этом наборе есть слова разных частей речи: 666 пар с существительными, 222 пары с глаголами, 111 пар с прилагательными. Недостатком здесь будет отсутствие пар с разными частями речи в паре, как в MEN.

  В наборе дополнительно есть следующие теги: часть речи, рейтинги конкретности обеих слов (concreteness rating), сила ассоциации между словами (The strength of free association from word1 to word2), стандартное отклонение всех оценок экспертов на этой паре (The standard deviation of annotator scores when rating this pair) — можно использовать для оценки уверенности в оценке близости.

Резюмируем результаты анализа в Таблице 2.

Название	Что оценивалось?	Число разметчиков	Распределение по частям речи
WordSim353	Relatedness	13 на 153 пары/16 на 200 пар	100% n
The MEN Test Collection	Relatedness	1 на каждую пару	81.5% n + 12.76% j + 5.73% v
SimLex-999	Similarity	500	66.6% n + 11.1% j + 22.2% v

Таблица 2: Особенности «золотых стандартов».

Выявление и исправление дефектов в «золотых стандартах»

Выше были упомянуты проблемы, связанные с неразличимостью similarity и relatedness. Эти проблемы являются следствием недостаточно точно сформулированных инструкций экспертам по разметке, их исправить мы не сможем.

Интересует вопрос, сохраняется ли в «золотых стандартах» свойство симметричности. Вполне естественно считать, что на паре (w_1, w_2) и на паре (w_2, w_1) мера близости должна совпадать. Проверим выполнение этого свойства в выбранных наборах данных.

- В WordSim353 — обнаружено 2 пары с симметричностью слов: (money, bank) и (tiger, tiger). Пару (tiger, tiger) создатели датасета считают дефектом, его в нем быть не должно. А на паре (money, bank) симметричность не сохраняется — оценки связанности отличаются. Кроме того, был обнаружен дубликат (money, cash).

- В MEN симметричных пар слов не обнаружено, дефектов с дублированием и одним и тем же словом в паре не найдено.

- В SimLex-999 найдена одна пара с симметричностью слов (strange, sly), при этом оценки связанности отличаются.

Продублируем результаты в Таблице 3.

Название	Симметричность	Дефекты
WordSim353	Нет; 'money-bank'	пара 'tiger-tiger'; дубликат 'money-cash'; несимметричность 'money-bank'
The MEN Test Collection	Да; отсутствуют зеркальные пары	Не обнаружено
SimLex-999	Нет; 'strange-sly'	Нет; 'strange-sly'

Таблица 3: Дефекты «золотых стандартов».

В случае WordSim353 и SimLex-999 можно удалить дефектные пары, и считать, что во всех трех наборах данных выполнено свойство симметричности (будут отсутствовать зеркальные пары).

Разделение «золотых стандартов» на обучение и контроль

Для дальнейшего пользования библиотекой будет полезным функционал разделения «золотого стандарта» на обучающую и контрольную (тестовую) выборки. В нашем случае, выборка представляет из себя набор пар слов и величин сходства между ними. Обычное случайное разрезание части пар на обучение и части на контроль не подойдет. Причины следующие:

- На исходном наборе данных есть некоторое распределение по частям речи, которое хотелось бы сохранить и в обучающей, и в контрольной выборке.
- Если «резать» по парам, теряется первоначальная логика: работа идет со словами, а не с парами. Хочется проверить потенциальную модель не только на парах, которые она видит в первый раз, но и на незнакомых для нее словах.

По этим причинам можно сформировать требования к ожидаемому разбиению на обучение и тест:

- На полученных обучающей и тестовой выборке должно сохраняться распределение частей речи, как и на исходных выборках.
- Множество слов «золотого стандарта» должно разбиться на три части: подмножество слов, которые встречаются только в обучающей выборке, подмножество слов, которые встречаются и в обучающей, и в контрольной, и подмножество слов, которое встречается только в контрольной выборке.

Причем эти условия должна солюдаться при произвольных долях размера обучающей и тестовой выборки относительно исходного «золотого стандарта», доля задается пользователем. Размер подмножества слов, которое встречается только в контроле, задается пользователем, как доля уникальных для контроля слов от размера множества всех слов.

Решать поставленную задачу будем следующим образом:

1. Сперва заполняем контрольную часть парами с уникальными для контроля словами. Сэмплируем из распределения частей речи на «золотом стандарте» некоторую часть речи. Случайно выбираем слово с такой частью речи. Это слово должно оказаться только в контроле, поэтому все пары, где встречается это слово, отправляем в контроль.

2. Теперь у нас есть некоторое заполнение контроля. Нам нужно добрать пар в контроль до нужного размера. Аналогично, сэмплируем часть речи, выбираем случайную пару с такой частью речи. Проверяем, что для каждого слова в паре есть другие пары, где оно встречаются (таким образом убеждаемся, что мы не добавим еще слов во множество уникальных для контроля слов). Отправляем пару в контроль.

3. Все оставшиеся пары отправляем в обучение.

Некоторые замечания насчет сэмплирования: очевидно, что в случае одинаковых частей речи в каждой паре выбранный нами метод разделения сохранит распределение частей речи в полученных датасетах. В случае же разных частей речи в паре, что встречается, например, в датасете MEN, при выборе слова, которое будет уникальным для контроля, мы забираем все слова, которые входят с ним в пару. Вообще говоря, распределение частей речи при таком сэмплировании может незначительно смещаться. Это место для потенциальных будущих улучшений.

Стандартизация данных

На данный момент важными тегами слов были приняты части речи. Поэтому в стандартизованном виде необходимо отразить непосредственно два слова в паре, части речи каждой из них, и меру сходства. Для хранения данных на диске выберем формат таблиц csv с столбцовым разделителем запятая. Каждая строка отвечает паре слов и записывается в формате: $word_1, word_2, label_1, label_2, value$.

Программная реализация

Для представления пары слов в программе написан класс $Wordpair$, для него в частности определена операция взятия хэша. За счет этого, объект пара слов может быть использован в качестве ключа в словаре. Таким образом, для хранения в памяти «золотого стандарта» выбрана структура данных словарь. Этой структуре в программе отвечает класс $Data$. Он же хранит метки частей речи каждого слова в паре. Для взаимодействия с этой структурой данных разработан соответствующий интерфейс, позволяющий добавить пару, удалить пару, вернуть все пары с заданным словом и т.д.

Непосредственно «золотому стандарту» отвечает класс $GoldenStandartDataset$. Сюда вынесен функционал загрузки набора данных в память с диска по заданному пути, запись стандарта в файл, а также дополнительный интересующий нас функционал. Подробнее об этом функционале:

- Поиск зеркальных пар, проверка на симметричность, поиск всех асимметричных пар.
- Вычисление распределения частей речи на наборе данных.
- Нахождение множества всех слов в наборе данных вместе с соответствующими им метками частей речи.
- Разбиение «золотого стандарта» на обучение и контроль с сохранением пропорций частей речи и с заданной долей уникальных для контроля слов.

Класс $GoldenStandartDataset$ считывает в память наборы данных только в стандартизированном виде, который обсуждался выше (аналогично с записью в файл). Для рассматриваемых в работе трех датасетов написаны соответствующие классы: $SimLex999Dataset$, $WordSim353Dataset$, $MENDataset$. Каждый такой класс является наследником общего класса, реализующего «золотой стандарт», расширяя функционал считывания в память из формата, который использовал создатель выбранного датасета.

Для непосредственного оценивания качества предоставленных пользователем векторных представлений слов предлагается использовать класс $EmbeddingEvaluator$. На данный момент в нем реализовано оценивание с помощью ранговой корреляции между близостями, посчитанными по представлениям, и близостями из «золотого стандарта». Для вычисления близостей по представлениям пользователь может подать свою функцию расстояния, по умолчанию же используется косинусное расстояние.

Результаты

В рамках выполнения задания преддипломной практики было сделано:

- Автор ознакомился с тематикой дистрибутивной семантики, изучил, как можно оценивать представления слов.
- Были выбраны «золотые стандарты», проведен анализ метода построения каждого из них, отмечены особенности и недостатки, проведено сравнение.
- В выбранных стандартах были выявлены дефекты, для упрощения процедуры поиска дефектов реализованы соответствующие программные функции. Где это возмжожно, дефекты устранены.
- Были видвинуты требования к «грамотному» разбиению наборов данных на обучающую и контрольную выборки. Был предложен способ достижения этих требований, соответствующая функция реализована.
- Предложен стандартный формат хранения «золотого стандарта», выбранные наборы данных приведены к этому формату. Построено фиксированное разбиение «золотых стандартов» на обучение и контроль.
- Выбрана структура данных для хранения «золотого стандарта», написан соответствующий класс. Для удобного взаимодействия с парами слов и «золотыми стандартами» написаны соответствующие классы. При разработке интерфейсов предполагалась возможность быстрого изменения логики одного компонента без необходимости изменять всю библиотеку.

- Выбран способ оценивания пришедших на вход пользовательских представлений. Работа с оцениванием вынесена в отдельный класс.

Источники

1. WordSim353: http://gabrilovich.com/resources/data/wordsim353/
2. WordSim353 с разделением similar и related: http://alfonseca.org/eng/research/wordsim353.html
3. The MEN Test Collection: https://staff.fnwi.uva.nl/e.bruni/MEN
4. SimLex-999: https://fh295.github.io/simlex.html