Визуальная интерпретация статического векторного пространства для текстов на русском языке

О.А. Сериков [1], Э.С. Клышинский [2], В.А. Ганеева [3]

Национальный исследовательский университет «Высшая школа экономики», ул. Мясницкая, д. 20, Москва, 101000, Россия
[1] ORCID: 0000-0002-3746-2642, srkvoa@gmail.com
[2] ORCID: 0000-0002-4020-488X, eklyshinsky@hse.ru
[3] ORCID: 0000-0002-9569-9197, vaganeeva@edu.hse.ru

Аннотация

С момента появления статических векторных представлений слов было известно, что в них работает задача аналогий. В ней утверждается, что можно найти такой вектор, который переносил бы одно слово в другое, заменяя при этом некоторый признак. Также было известно, что задача аналогий не всегда решается качественно, в связи с чем возникла задача исследования свойств векторных семантических пространств. В данной работе мы вводим метод визуальной интерпретации такого пространства. Основой метода является использование тематических коллекций слов, разделение векторного пространства при помощи метода LSA и визуализация результатов с использованием тепловых карт. В ходе экспериментов мы обнаружили, что векторные пространства могут быть интерпретированы не только на локальном, но и на глобальном уровне. Разделение пространства на части при этом зависит от набора текстов, на котором проводилось создание этого пространства. Метод оказался пригоден для выделения нескольких верхних уровней, так как при увеличении глубины анализа количество слов в группе сокращается экспоненциально.

Ключевые слова

Статическое векторное пространство, визуальная интерпретация, LSA.

1. Введение

Как это было показано в работе [1], семантические векторные пространства обладают возможностью решения задачи аналогии. Формально, задачу поиска аналогий в векторном пространстве, где каждому слову соответствует точка в многомерном пространстве, можно поставить следующим образом. Пусть $v_{a'}$ и v_a — векторы, соответствующие словам a' и a. В этом случае, разница между векторами $v_{a'} - v_a$ будет показывать семантическое отношение между ними. Тогда для известного слова b можно найти слово y, такое, что между ними будет иметься такая же аналогия (если слово y существует): $y = v_b + v_{a'} - v_a$.

Как показали исследования [2] и [3], подобные аналогии могут быть найдены не для всех слов. Это связано, например, с тем, что слова обладают несколько разными значениями. Так, в примере «король-мужчина+женщина=королева», королева не всегда является полновластным монархом, определяющим внешнюю и внутреннюю политику, а может быть лишь супругой короля с иной зоной ответственности. С другой стороны, такие работы как [4] показывают, что задача аналогии может быть решена с довольно высокой точностью.

Задачу аналогий можно переформулировать следующим образом. Пусть дано статическое векторное пространство, в котором определен некоторый вектор m, соединяющий две области пространства. Пусть даны слова a', a, b' и b, для которых справедлива некоторая общая аналогия (например, разные названия для одной должности, занимаемой мужчиной или женщиной, или отношение между государством и его столицей). Тогда для соответствующих векторов слов соблюдается следующее соотношение:

$$v_{a'} - v_a \approx v_{b'} - v_b \approx m. \quad (1)$$

Если $v_{a'}$ и $v_{b'}$ являются соседями, то есть принадлежат одной небольшой области пространства, то из соотношения (1) следует, что v_a и v_b также являются соседями. Из этого, в свою очередь

следует, что вектор *m* соединяет две области пространства, к которым относятся две группы слов, обладающие общими признаками внутри групп. Более того, слова из разных групп противопоставляются друг другу по некоторому признаку (для наших примеров это «мужчина-женщина» или «государство-столица»). Далее встает вопрос о том, насколько компактными являются данные области, и принадлежат ли им только те слова, которые обладают заданным признаком.

Практика показывает, что на локальном уровне подобное разделение соблюдается в большинстве случаев. Так, на Рисунке 1 показана область пространства, заполненная названиями трав, цветов, кустарников и деревьев, полученная на модели Araneum_upos_skipgram (http://rusvectores.org/ru/models/). Некоторое смешение слов может быть объяснено тем, что для получения рисунка использовался метод UMAP, дающий искажения во взаимном расположении точек. При увеличении числа измерений разделение между растениями улучшается, но ухудшается возможность визуализации. Как видно из рисунка, растения расположены по группам: съедобные противопоставлены несъедобным, приправы противопоставлены овощам и другим сельскохозяйственным растениям, деревья с добываемой древесиной противопоставляются плодоносным.

В данном исследовании мы проверяем гипотезу о том, что подобное разделение может быть автоматически обнаружено не только локально, то есть для малых групп слов, но и на верхнем уровне модели, то есть для больших групп слов, составляющих модель в целом или ее значимых фрагментов. Для проверки этой гипотезы мы разработали метод автоматического разделения пространства на семантически связанные фрагменты и визуализации результатов такого разделения.

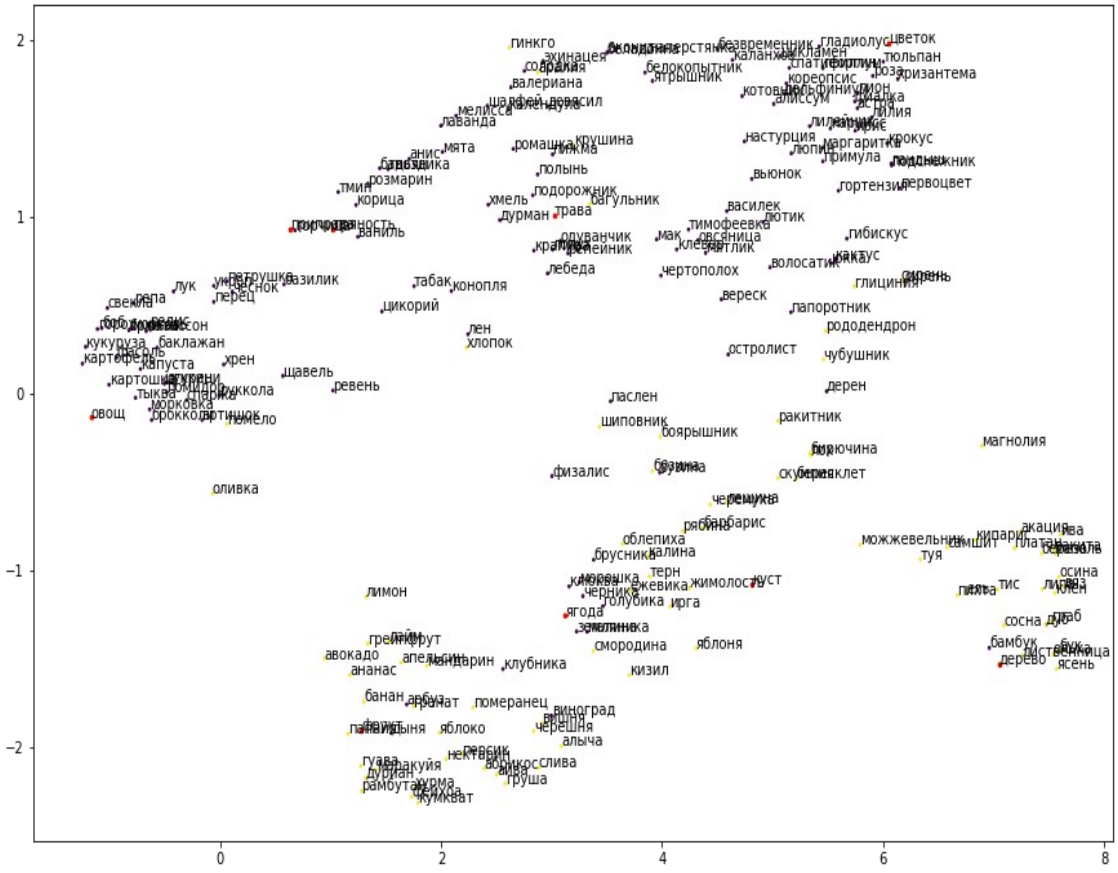

Рисунок 1 – Разделение слов, обозначающих растения, в статическом векторном пространстве

2. Существующие методы анализа семантического пространства

В последние годы для интерпретации семантических свойств моделей языка используется пробинг — подход при котором модель воспринимается как чёрный ящик, при этом

исследуется её реакция на некоторые входные стимулы и их изменение. Подобным образом может исследоваться не только семантические, но и грамматические свойства языковой модели. Пробинговые исследования можно структурировать как реализующие три последовательных этапа анализа моделей [5]: бихевиоральный, диагностический и инвазивный. В случае бихевиорального пробинга проверяется поведение модели с точки зрения изменения грамматических характеристик, например, будет ли нейросеть, генерирующая тексты, выдавать корректный результат при изменении грамматической категории числа в тексте-«затравке». Диагностический пробинг выявляет зависимость между грамматической характеристикой и поведением модели. Для этого в них измеряется линейная корреляция между векторным представлением текста и грамматической характеристикой. Так, например, на Рисунке 2 (взят из [6]) показана вычисленная зависимость между точностью решения определенной задачи (приведены по горизонтали: определение сходства текстов, вывод на тексте, задачи классификации слов и др.) и наличием в текстах определенной грамматической характеристики (приведены по вертикали: длина предложения, глубина дерева зависимости, количеством субъектов и объектов в предложениях и др.). Наконец, инвазивный пробинг использует внесение шума в текст с тем, чтобы понять, насколько он будет «понятен» модели, то есть насколько изменится качество анализа или генерации текста. Так, в работе [7] исследуется влияние ошибок в тексте на точность работы итоговой модели (см, например, Рисунок 3). Подобные исследования помогают понять, насколько чувствительна модель к определенным параметрам, и использует ли она их вообще.

Подобные исследования обычно проводятся на контекстуализированных языковых моделях. Основным отличием контекстуализированных и статических моделей является учет контекста употребления слов. В случае статических моделей вектор присваивается собственно слову, без учета контекста его употребления. В этом случае не учитывается семантическая или грамматическая многозначность слов. Для контекстуализированных моделей вектор слова может быть получен только с учетом контекста, то есть окружающих его слов. В связи с этим одному и тому же слову, употребляемому в кардинально разных контекстах, будут соответствовать разные вектора. В связи с этим встает проблема объединения различных смыслов одного слова. Заметим также, что слова, оказавшиеся рядом в тексте, могут получить близкие векторы, даже несмотря на то, что их предметная область или значение различны.

Еще одним направлением, активно развивающимся в последнее время, является исследование исторического сдвига смысловой нагруженности понятий. На начальном этапе исследователи обратили внимание, что с течением времени слова имеют тенденцию менять соседство, переходя из одной области в другую. Так, например в диссертации [8] предложен метод определения подобного семантического дрейфа относительно других слов. На Рисунке 4 на примере слова *monumental* наглядно показано, как данное слово переходит из области архитектуры в разговорную речь. Аналогичные исследования проводятся на данный момент для всех крупных языков.

Следующим шагом стало исследование взаимного расположения слов в зависимости от их окрашенности по некоторому признаку. Так, в работе [9] было показано, что различные виды спорта ассоциируются в текстах с разным уровнем достатка (от спортивного туризма и бокса, ассоциирующихся с бедностью, до тенниса и гольфа, ассоциирующихся с достатком).

Исходя из проведенного обзора, можно утверждать, что в семантическом пространстве, задаваемом при помощи моделей Word2Vec, существуют выделенные направления, задающие изменение некоторых признаков. Так, задача аналогии показывает, что существуют некоторые признаки, накладывающиеся на изменение смысла слов, относящихся к разным группам. Работа [9] показывает, что подобные направления могут быть интерпретированы понятным для человека образом.

В данной работе мы продолжаем исследования в области интерпретируемости статических семантически пространств. В предыдущих работах интерпретируемость направлений в семантическом пространстве проверялась на локальном уровне: уровне отдельных слов и их групп. Здесь мы проверяем гипотезу о том, что семантическое пространство в целом также имеет некоторые глобальные направления, разделяющий все слова в целом на связанные группы, причем размер этих групп сопоставим с размером словаря системы в целом.

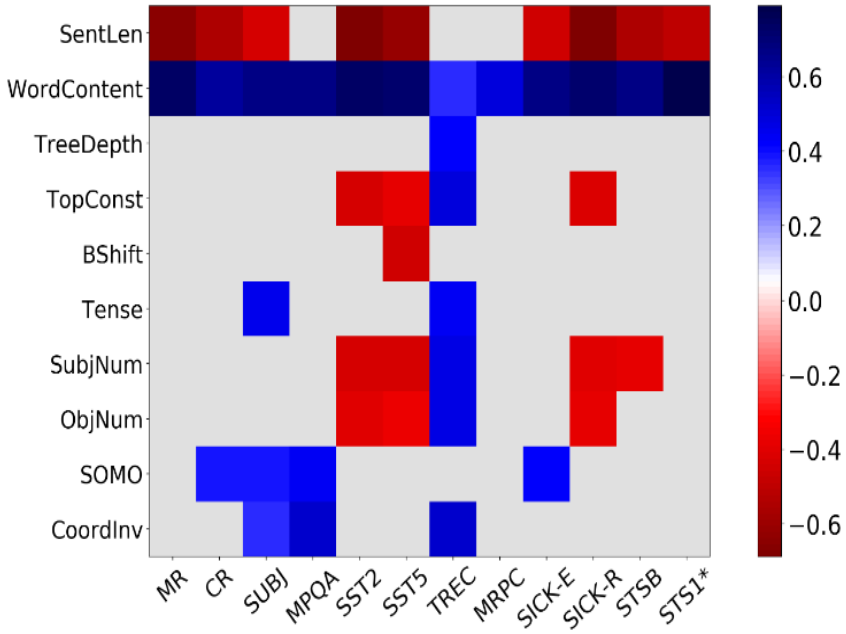

Рисунок 2 – Матрица корреляции решения различных задач в зависимости от представленности в тексте грамматических признаков (рисунок взят из [6]).

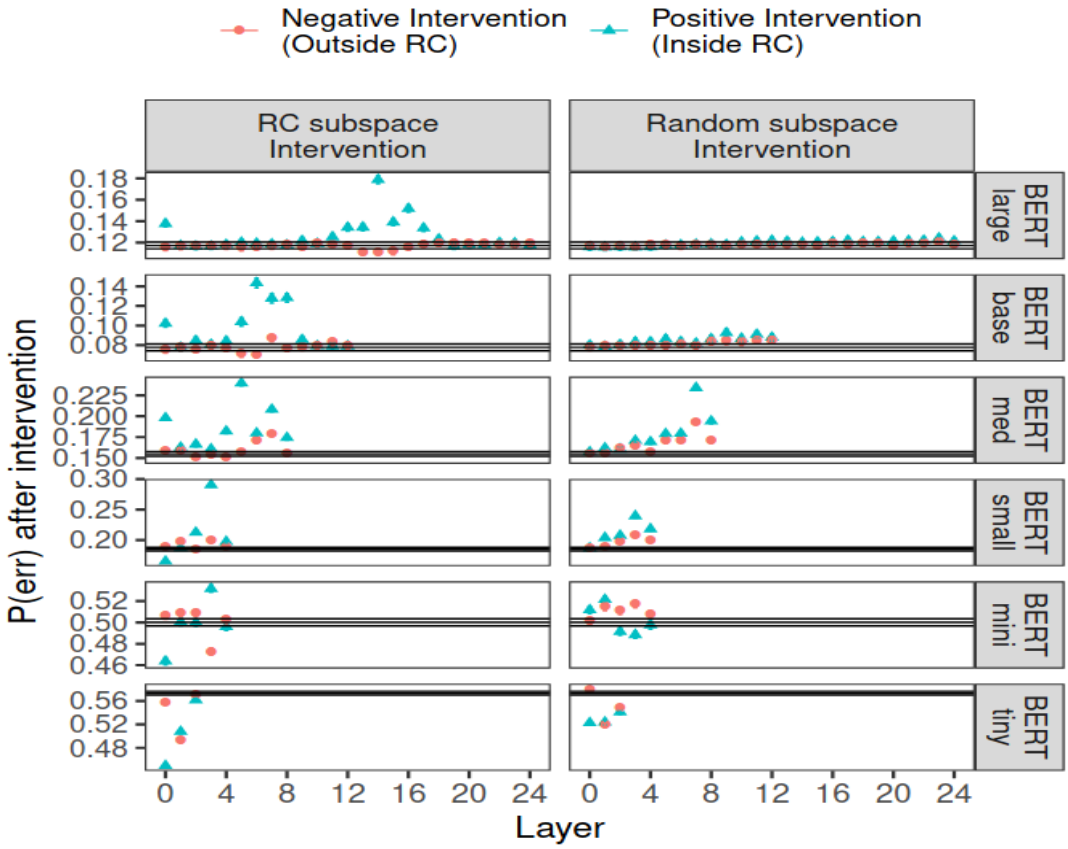

Рисунок 3 – Изменение значений коэффициентов на слоях нейронной сети в зависимости от внесения ошибки в согласование слов в тексте (рисунок взят из [7]).

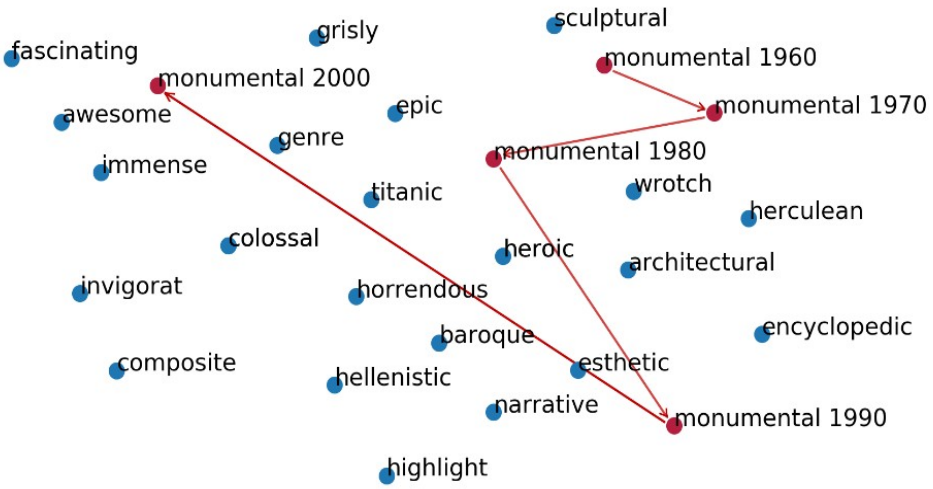

Рисунок 4 – Изменение смысла употребления слова *monumental* со временем (рисунок взят из [8]).

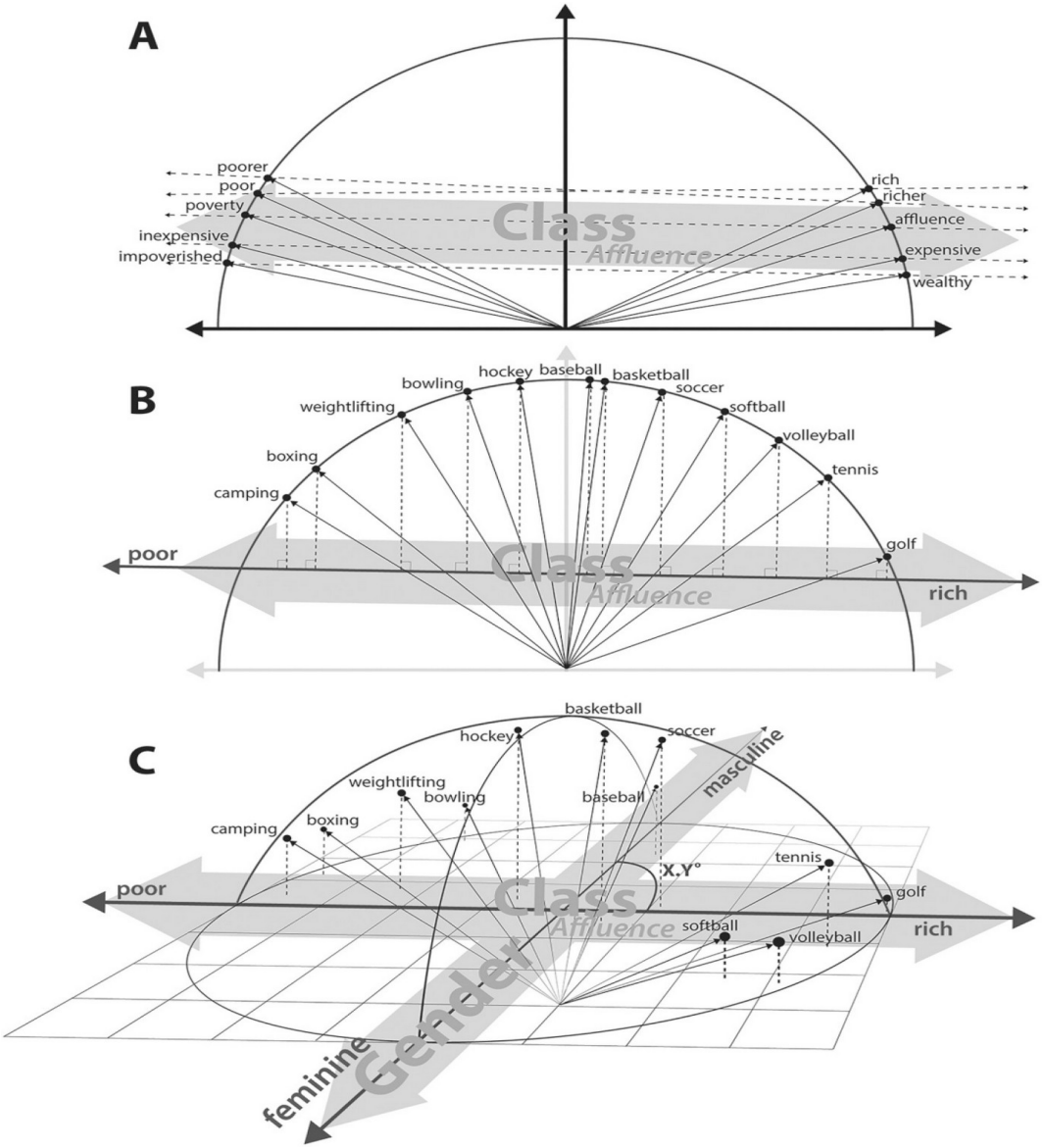

Рисунок 5 – Взаимное расположение слов в зависимости от их окрашенности по определенному признаку (рисунок взят из [9]).

3. Метод анализа статического семантического пространства

Для выделения семантически связанных групп слов обычно используется метод тематического моделирования [10]. Однако, данный метод обладает такими недостатками, как низкое количество уровней разбиения слов по тематикам и недостаточная интерпретируемость результатов. В связи с этим, мы решили использовать непосредственно метод, положенный в основу тематического моделирования — латентно-семантический анализ [11]. С одной стороны, он позволяет создать некоторое латентное пространство, лучше отображающее разделение признаков исходных данных. С другой стороны, в основе данного метода лежит метод сингулярного разложения векторов, осуществляющий поворот и масштабирование исходного пространства по осям, вдоль которых наблюдается наибольшая относительная дисперсия. Последняя позволяет лучше разделить слова между собой.

Для исследования мы использовали векторные представления из статических моделей Word2Vec. За счет этого имеется возможность зафиксировать семантическое пространство и исследовать его свойства. Здесь мы специально избегаем сложностей работы с такими контекстуализированными моделями, как BERT, так как для работы с ними необходимо тщательно подбирать коллекцию текстов и усреднять полученные вектора. Интерпретация же статической модели FastText (равно как и моделей BERT) затруднительна, так как она хранит только фрагменты слов.

Для разделения слов на тематические группы мы использовали следующий алгоритм. Пусть дан словарь слов $d = \{w_i\}$, по которым можно получить вектора Word2Vec и сформировать из них матрицу E =Word2Vec(d). Пусть счетчик числа пройденных шагов n = 1. К матрице E и словарю d можно применить следующий алгоритм разделения слов.

1. При помощи алгоритма LSA для матрицы E получим матрицу векторов в преобразованном пространстве R = LSA(E).
2. Выделим из матрицы R вектор номер n: $r = R_n$.
3. Отсортируем все слова по их значениям в r: $w' = argsort(w, r)$.
4. Разделим w' на три части в соответствии со значениями r: $d' = <d^-, d^0, d^+>$. Положим n = n + 1. До достижения необходимого уровня вложенности рекурсивно применим алгоритм разделения слов по главной оси к словарям d^- и d^+ и соответствующим им матрицам E^- и E^+.

Итогом работы алгоритма будет иерархия осей (векторов), задающих принципы разделения пространства. Заметим, что иерархия представляет собою дерево, в котором сверху идут более общие признаки. Подобная структура связана с тем, что слова, принадлежащие к разным классам, не имеющим общего класса более высокого уровня, будут обладать разным набором признаков. Так, например, абстрактные понятия не обладают протяженностью и другими признаками физических объектов.

На каждом уровне разделения мы анализировали только слова, попавшие в словари d^- и d^+, пропуская при этом словарь d^0. За счет этого на каждом новом уровне иерархии количество столбцов увеличивается в два раза по сравнению с предыдущим.

Мы извлекли из Wiktionary тематические списки слов из следующих категорий: геология, география, минералы, растения, оружие, искусства, филология, философия, информатика, архитектура, фортификация, геологические эпохи, политика, профессии, ранги, занятия, имена мужские и женские, устаревшие слова, российские города и реки, одушевлённые существительные. Эти категории использовались для контроля разделения словарей по тематикам. Для более простой визуализации результатов, мы специально подбирали слова из лексики, относящейся как к далеким друг от друга тематикам (устаревшая лексика против современной, гуманитарные и технические области знаний), так и к близким (такие контактирующие области как геология и география). Наличие списков слов по категориям позволяет визуализировать их разделение на разных уровнях.

Метод визуального анализа полученного разделения заключается в следующем. Для каждого уровня строится тепловая карта, представляющая собой таблицу, каждая ячейка которой окрашена с использованием градиента (в интервале между минимальным и максимальным значениями). В этой таблице столбец сопоставлен со словарем соответствующего уровня, а строка – с одной из категорий. Для каждой строки считается

пересечение категории со словарями, в ячейке показана доля слов из пересечения категории и словаря относительно пересечения категории со всеми словарями на данном уровне разделения. Для улучшения визуального восприятия мы использовали следующую формулу для вычисления отображаемого значения в ячейке v_{ij}, принадлежащей словарю i и категории j:

$$v_{ij} = \frac{\log\left(1 + words_{ij}\right)}{1 + words_j}, \quad (2)$$

где $words_{ij}$ — количество слов в словаре i из категории j, а $words_j$ — количество слов в категории j. Подобное отображение позволяет повысить контрастность изображения, так как при использовании относительных значений ряд особенностей в распределении скрадывается.

Тепловая карта позволяет визуализировать разделение категорий по частям пространства и носит иерархический характер. Так, на первом уровне пространство делится на две части, далее каждая из частей снова делится пополам; то есть, разделение на столбцы можно рассматривать как двоичное дерево. Но так как мы не контролируем направление осей, получаемых в результате LSA, порядок следования словарей может меняться. То есть мы не можем утверждать, что соседство ячеек внутри строки означает их соседство в исходном пространстве.

4. Визуальный анализ разделения слов на верхних уровнях иерархии

В данном разделе мы покажем результаты для модели Word2Vec, которую мы обучили на текстах научных статей, относящихся к таким разделам как архитектура, искусства, автоматизация, геология, история, культура, лингвистика, литература. Помимо этой модели, мы использовали несколько, взятых с сайта RusVectores, однако их анализ выходит за рамки данной статьи. Заметим только, что метод визуального анализа хорошо показал себя и на них.

На первом разделении была получена тепловая карта, показанная на Рисунке 6. По ней видно, что в одну группу попали слова, относящиеся к геологии, названиям минералов и городов, частично — одушевленные объекты и термины из филологии, философии и политики.

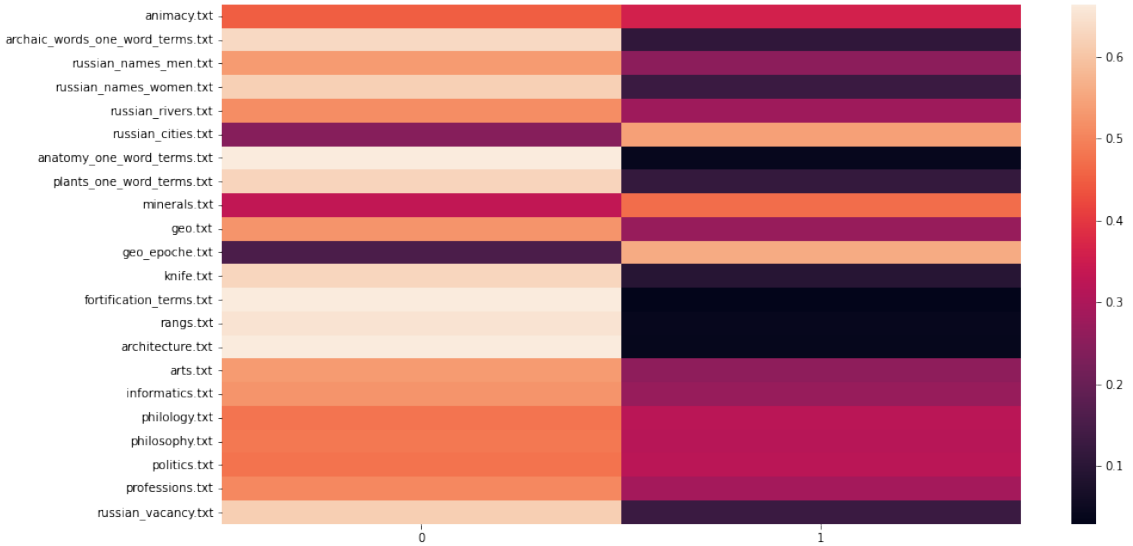

Рисунок 6 – Разделение слов на первом уровне иерархии

Более подробный анализ списка слов показал, что имеет место противопоставление бытового и научного дискурсов. Максимальными значениями по оси обладали слова, используемые в повседневном общении, минимальными – фамилии учёных и авторов статей, названия институтов, университетов и других образовательных и научных учреждений, города их расположения, сложные научные термины (например, «интертекстуальность» или «линеаризация»). Несмотря на то, что различие между этими группами можно описать как разницу между более научным и более бытовым дискурсами, это не означает, что в группу более бытового не могут попасть научные термины. Однако сложность терминов из первой

группы будет меньше, чем слов из второй. Например, «стих» будет в группе более бытового, а «дольник» или «акростих» – в группе более научного. Аналогично, в области информационных технологий на одной стороне оси находились такие слова, как «регистр», «код», «аргумент», «почта», «ядро», «модуль», «компьютер», «буфер», «сценарий», «контейнер», «субъект», «протокол», а на другой стороне — «кэширование», «транслятор», «репликация», «октет», «инкапсуляция», «кодирование», «трекер», «эмуляция», «битрейт», «профайл», «квантор», «селектор». Заметим, что первые слова часто встречаются в новостях или в разговоре, тогда как вторые в основном присущи специальным статьям или обсуждению профессиональных тем.

К третьему уровню разделения тематики сузились. На Рисунке 7 группа 0 – содержит слова из списков занятий, профессий, политики и рангов-должностей. Группа 1 содержит термины философии и филологии. Группа 2 содержит названия растений, минералов, оружия, термины географии, фортификации и архитектуры – объекты, которые именно под этими названиями чаще фигурируют в статьях, чем в бытовом дискурсе. Для групп 4 и 5 в списках нашлось мало слов из-за узкой специализации, группа 6 – имена, группа 7 – города и организации. Здесь тоже видно, что конкретные фамилии исследователей и организации принадлежат к научному дискурсу. В целом, к третьему уровню выделились темы, представленные в Таблице 1.

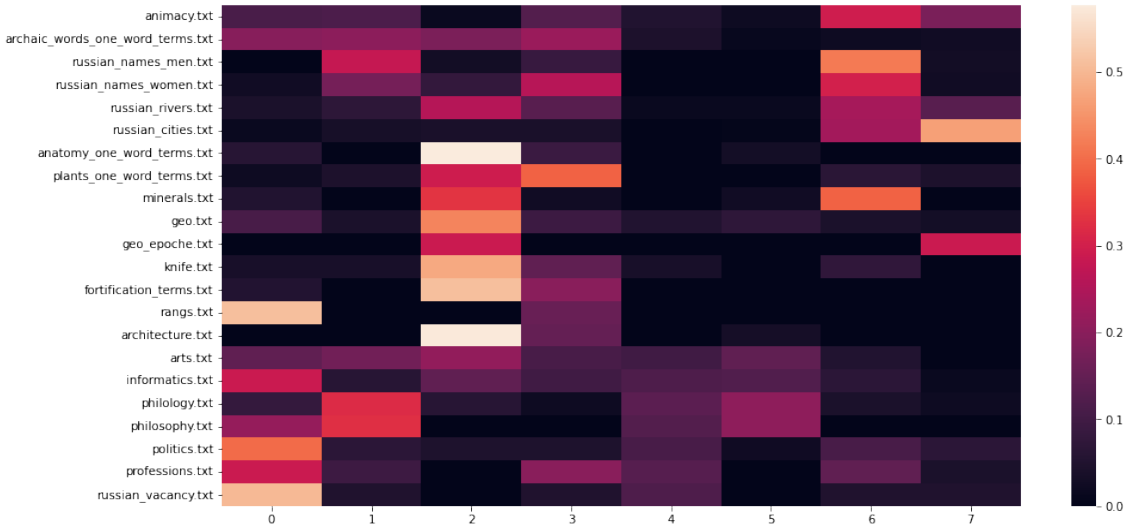

Рисунок 7 – Разделение слов на третьем уровне иерархии

Таблица 1 – Тематики, на которые разделились слова

бытовой дискурс				научный дискурс			
абстрактное		предметы		научные термины		места и люди	
общество и политика	книги и духовный мир	спец. термины	бытовые предметы	внеш. научное	внутр. научное	фамилии и имена учёных	места и организации

Можно было бы предположить, что модель, обученная на научных статьях, начнёт с разделения между собой научных сфер и дисциплин, однако первое разделение показало, что в самом начале эти оси также показывают более универсальные признаки. На практике, разделение на отрасли происходит только после выделения методов исследования и объектов исследования, так как их описания в статьях из разных предметных областей строятся с использованием сходных конструкций. На третьем уровне группа терминов, отвечающих за научный дискурс, разделилась на специальности, явления и процессы, происходящие с науками или научными дисциплинами (например, «изыскание», «совершенствование», «специальность», «освоение», «ознакомление», «диссертация», «методология»), и на процессы и явления, происходящие с объектами науки (например, «аналитичность», «стереотипность», «субъективность», «предшествование», «обязательность», «разорванность», «асимметрия», «локальность»).

На шестом уровне разделения выделяются группы размером около 30 слов (см. Рисунок 8). Общие тенденции разделения сохраняются, однако вероятность попадания всех слов категории в один словарь весьма мала. Здесь же начинают ощущаться недостатки предложенного метода. В связи с тем, что фильтрация по частоте не проводилась, в категориях «ранги» и «геологические» осталось лишь по одному термину, которые показывают максимальное значение на тепловой шкале. Более того, сами эти слова («советник» и «пермь») относятся скорее к выделенным словарям, чем к своим категориям (хотя к категориям они, безусловно, тоже имеют прямое отношение).

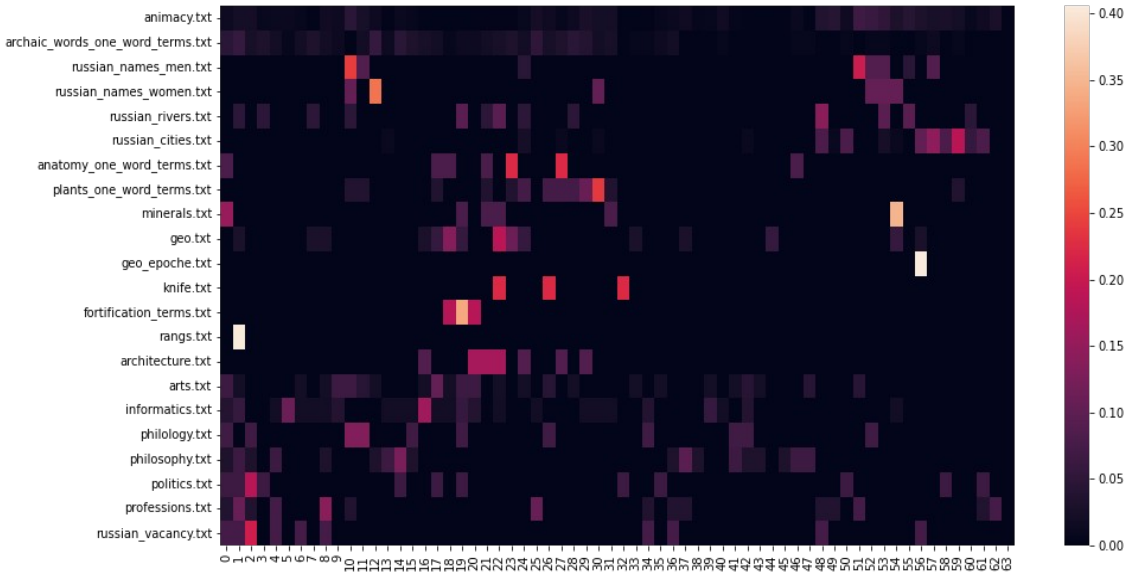

Рисунок 8 – Разделение слов на шестом уровне иерархии

Предложенный метод показал, что при смене векторной модели меняется и ее интерпретация., однако во всех моделях можно обнаружить разделение на на более бытовое и более специальное, материальное-нематериальное, абстрактное-конкретное. Заметим, что один и тот же вида разделения может быть встречен разной глубине разделения, то есть единого разделения для всех моделей построить не получается.

Также заметим, что получающееся семантическое пространство не является метрическим в общем смысле этого термина. При первичном разделении по некоторой оси

5. Заключение

В данной работе мы проверили гипотезу об интерпретируемости семантического пространства, полученного при помощи статических моделей Word2Vec, не только на локальном уровне, но и на уровне всей модели в целом. Наши исследования показали, что начальный уровень интерпретации зависит от того, на каких текстах обучалась модель. Так, например, при использовании беллетристики разных эпох модель выделяет области абстрактного и конкретного, а на втором уровне противопоставляет духовное прикладному, современное — архаичному. Для текстов, взятых из сети Интернет, сперва разделяется социальное и организационное, а затем абстрактное противопоставляется конкретному, а технология — управлению.

Для анализа мы использовали метод визуализации принадлежности слов тому или иному разделению. Метод основан на применении тепловой карты к долям слов предметной области (категории), отнесенной к той или иной области пространства (словарю). Метод хорошо показал себя на практике, однако стали очевидны некоторые недостатки. Так, например, на глубоких уровнях анализа количество слов экспоненциально падает. Как следствие, падает и вероятность найти слова из фиксированного списка терминов предметной области (категории) в маленькой зоне пространства. Более того, сами термины категорий оказываются многозначными и находятся сразу в нескольких зонах, относящихся к разным предметным

областям. Более строгий выбор слов, входящих в категории, представляет собой отдельную лингвистическую задачу.

Анализ нескольких языковых моделей, обученных на текстах разных предметных областей, показал, что состав выделяемых осей зависит от использованной лексики. При этом некоторые оси выделяются для всех моделей, хотя и находятся на разном уровне иерархии. Это позволяет говорить о некоторой универсальности самих осей, но не их взаимного расположения.

Еще одной проблемой является выбор границы между словарями, так как строгое назначение порогового значения разделяет термины одной предметной области, находящиеся в семантическом пространстве рядом. Решением может быть применение кластеризации и отнесение к словарю всего кластера целиком.

Наконец, мы проводили анализ только слов, оказавшихся на периферии, тогда как большинство слов попадают в центр, где плотность расположения терминов значительно выше. Все эти проблемы могут быть решены в ходе дальнейших исследований.

6. Список источников

[1] Mikolov T., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality // In Proc. of Neural Information Processing Systems 27: 27th Annual Conference on Neural Information Processing Systems, 2013. P.3111-3119

[2] Korogodina O., Karpik O., Klyshinsky E. 2020. Evaluation of Vector Transformations for Russian Word2Vec and FastText Embeddings // Conference on Computer Graphics and Machine Vision: GraphiCon 2020. 2020. V. 2744. P. paper18-1 – paper18-12.

[3] Korogodina O., Koulichenko V., Karpik O., Klyshinsky E. 2021. Evaluation of Vector Transformations for Russian Static and Contextualized Embeddings // Conference on Computer Graphics and Machine Vision: GraphiCon 2021. 2021. V. 3027. P. 349-357.

[4] B. Wang, A. Wang, F. Chen, Y. Wang, J. Kou, Evaluating word embedding models: methods and experimental results // In Proc. of APSIPA Transactions on Signal and Information Processing, 2019, 8. doi: 10.1017/ATSIP.2019.12

[5] Lasri K., Pimentel T., Lenci A., Poibeau T., Cotterell R. Probing for the Usage of Grammatical Number / In Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, V. 1, pp. 8818–8831.

[6] Conneau A. et al. What you can cram into a single vector: Probing sentence embeddings for linguistic properties [Электронный ресурс]: arXiv preprint arXiv:1805.01070. – 2018. URL: https://arxiv.org/abs/1805.01070 (дата обращения 01.10.2022).

[7] Ravfogel S. et al. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction [Электронный ресурс]: arXiv preprint arXiv:2105.06965. – 2021. URL: https://arxiv.org/abs/2105.06965 (дата обращения 01.10.2022).

[8] Kutuzov A. Distributional word embeddings in modeling diachronic semantic change // Doctoral Thesis, University of Oslo, [Электронный ресурс]: https://www.duo.uio.no/bitstream/handle/10852/81045/1/Kutuzov-Thesis.pdf (дата обращения 01.10.2022).

[9] Kozlowski A., Taddy M., Evansa J. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. American Sociological Review, 2017, pp. 905-949.

[10] Воронцов К. В., Потапенко А. А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — Вып. 13 (20). — М: Изд-во РГГУ, 2014. — С. 676-687.

[11] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. Indexing by latent semantic analysis // Journal of the American Society for Information Science. 1990. V. 41, Iss. 6. P. 391-407.