Метод визуальной интерпретации статического векторного пространства Word2Vec

Э.С. Клышинский ¹, В.А. Ганеева ²

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.

A Method for Visual Interpretation of Word2Vec Static Vector Space

E.S. Klyshinskiy ¹, V.A. Ganeeva ²

Abstract

Starting its creation, static vector word models was able solving the task of finding word analogies. This task finds a vector of parallel transition that reflects a word into another one changing a property of a former word. Since the task of finding word analogies lacks of precision, researchers are trying to investigate peculiarities of vector models and corresponding semantic spaces. In this paper, we introduce a new method of visual interpretation of such models. Our method uses thematic word collections, separation of semantic space into word groups using LSA method, and visualization of results using heatmap. We found out that vector semantic space can be interpreted not only on a local level, but on the global one as well. In the former case, the resulting grouping of words depends on the text collection used for the model training. Our method is applicable for finding several top layers because of exponential decreasing of the size of a group.

Keywords

Static vector space, visual interpretation, LSA.

ORCID: 0000-0002-4020-488X (Э.С. Клышинский); 0000-0002-9569-9197 (В.А. Ганеева)



© 2022 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹ Институт прикладной математики им. М.В. Келдыша, Миусская пл., 4, Москва, 125047, Россия

² Национальный исследовательский университет «Высшая школа экономики», ул. Мясницкая, 20, Москва, 101000, Россия

¹ Keldysh Institute of Applied Mathematics RAS, Miusskaya sq., 4, Moscow, 125047, Russia

² National Research University "Higher School of Economics", Myasnitskaya st., 20, Moscow, 101000, Russia

ГрафиКон 2022: 32-я Международная конференция по компьютерной графике и машинному зрению, 19-22 сентября 2022 г., Рязанский государственный радиотехнический университет им. В.Ф. Уткина, Рязань, Россия EMAIL: klyshinsky@hse.ru (Э.С. Клышинский); vaganeeva@edu.hse.ru (В.А. Ганеева)

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:

$$v_{a\prime} - v_a \approx v_{b\prime} - v_b \approx m. \tag{1}$$

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 «мужчинаженщина» 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_skipgram². 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 противостоят 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.

2. Метод анализа статического семантического пространства

Для выделения семантически связанных групп слов обычно используется метод тематического моделирования [5]. Однако, данный метод обладает такими недостатками, как низкое количество уровней разбиения слов по тематикам и недостаточная интерпретируемость

1

² http://rusvectores.org/ru/models/

результатов. В связи с этим, мы решили использовать непосредственно метод, положенный в основу тематического моделирования — латентно-семантический анализ [6]. 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.

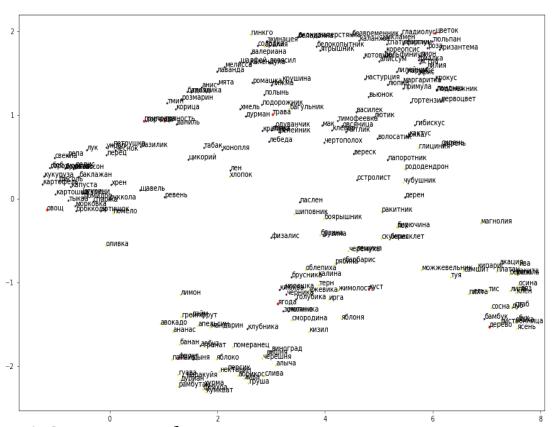


Рисунок 1 – Разделение слов, обозначающих растения, в статическом векторном пространстве

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: $d' = \langle d', d^0, d^+ \rangle$. 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: $v_{ij} = \frac{log(1+words_{ij})}{1+words_{j}},$

$$v_{ij} = \frac{log(1 + words_{ij})}{1 + words_{j}},\tag{2}$$

где $words_{ij}$ — количество слов в словаре i из категории j, а $words_{j}$ — количество слов в категории ј. 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.

3. 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 2. 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.

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 сложные научные термины (например, «интертекстуальность» расположения, «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.

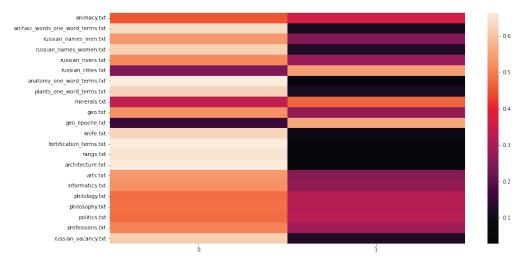


Рисунок 2 – Разделение слов на первом уровне иерархии

К третьему уровню разделения тематики сузились. На рисунке 3 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 группа 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.

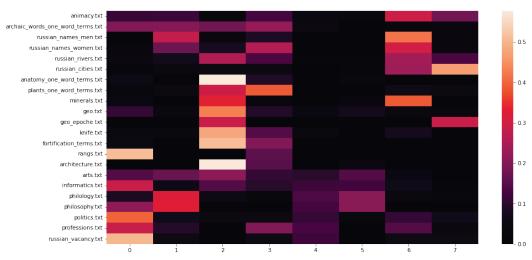


Рисунок 3 — Разделение слов на третьем уровне иерархии

Таблица 1 – Тематики, на которые разделились слова

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

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 4). 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).

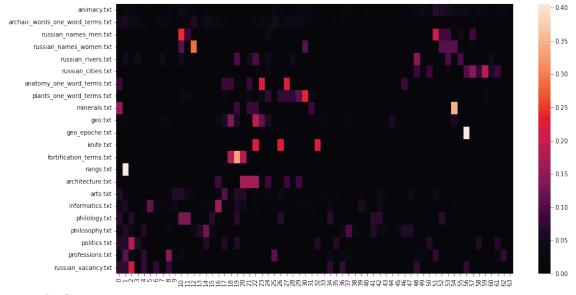


Рисунок 4 — Разделение слов на шестом уровне иерархии

4. 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.

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.

5. 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] 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.
- [6] 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.