

ПОРОЖДЕНИЕ ИНТЕРПРЕТИРУЕМЫХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ НА ОСНОВЕ СУПЕРПОНЯТИЙ

М.М. Тихомиров (M.M. Tikhomirov)^{a,b*}, Н.В. Лукашевич (N.V. Loukachevitch)^{a,b}**

^a МГУ имени М.В.Ломоносова, 119991, Российская Федерация, Москва, Ленинские горы, д. 1

^b Центр доверенного искусственного интеллекта, ИСП РАН, 109004, Российская Федерация, Москва, ул. А. Солженицына, д. 25

* e-mail: tikhomirov.mm@gmail.com

** e-mail: louk_nat@mail.ru

В данной работе представлен подход по созданию интерпретируемых векторных представлений слов, в которых каждая компонента вектора соответствует некоторой интерпретируемой семантической категории. Для получения таких категорий используется лексико-семантический ресурс в виде семантической сети RuWordNet, а также представительный корпус русскоязычных текстов для порождения векторных представлений. Полученные интерпретируемые векторные представления были протестированы на возможность отображать разные модели в одно векторное пространство.

Ключевые слова

векторные представления слов, семантические категории, тезаурус, семантическая близость

Введение

В настоящее время основной формой представления слов и других языковых единиц являются представления в виде векторов низкой размерности (эмбеддингов), получаемых за

счет обработки больших текстовых корпусов [6, 15]. При этом сами размерности векторов не являются интерпретируемыми – невозможно сказать, что означает та или иная размерность в векторе. Векторные представления формируются при анализе контекстов в корпусе, при этом в процессе обучения присутствует элемент случайности, в первую очередь при инициализации модели. Повторное вычисление векторных представлений на том же корпусе, векторные представления, рассчитанные на другом корпусе, не могут быть сопоставлены между собой, требуют специальных процедур построения сопоставлений (преобразований) между векторными представлениями.

Проблема сопоставления разных векторных представлений, посчитанных на разных корпусах одного языка или разных языков, возникает в таких задачах, как кросс-языковые переносы моделей [4, 14], переносы моделей между различными областями (domain transfer) [21], сравнение векторных представлений, обученных на текстовых корпусах, относящихся к различным временным интервалам (например, для изучения изменения значений слов).

В экспериментах разных авторов было показано, что векторные пространства содержат так называемые хабы (hubs), которые являются векторами, близкими к большому количеству других векторов в пространстве [18]. Эта проблема проявляется в векторных пространствах слов в виде слов, которые имеют высокое косинусное сходство с большим количеством других слов [19].

В данном исследовании мы рассматриваем проблемы хабов с другой стороны – можно ли выделить семантически интерпретируемые хабы в векторном пространстве, как некоторые существенные семантические категории в коллекции текстов, и далее развернуть векторное пространство с учетом этих хабов, сделав таким образом измерения векторного пространства интерпретируемыми.

Мы исследуем подход получения семантически интерпретируемых категорий для векторных представлений на основе привлечения лексико-семантических ресурсов в виде семантических сетей типа WordNet [7]. В данном исследовании мы проводим эксперименты

для русского языка, и в качестве таких хабов мы используем семантические классы (суперпонятия) из семантического ресурса русского языка RuWordNet. [13]. Отбор суперпонятий производится на основе представительного текстового корпуса, что позволяет создать адаптивный набор семантических категорий (суперпонятий) для задачи и/или предметной области.

1. Близкие работы

В теоретической и компьютерной семантике есть понятие семантического класса, т.е. совокупности слов, которые обладают похожими семантическими признаками, например живые существа, предметы, действия и др. Разные семантические теории и модели предлагают свои наборы семантических классов [3, 9], которые используются, например, при синтактико-семантическом анализе, для построения семантической структуры предложения и др.

При использовании в автоматическом анализе текстов ресурсов в виде семантических сетей, например, WordNet [7] такие семантические классы образуются естественным образом – за счет понятий (синсетов) на верху иерархии семантической сети (суперпонятий, supersenses). В WordNet такой список суперпонятий включает 26 синсетов существительных, например, ANIMAL (животное), PERSON (человек), FEELING (чувство) и 15 синсетов глаголов – COMMUNICATION (общение), MOTION (движение), COGNITION (познание).

В работе [8] авторы создают векторные представления для суперпонятий WordNet. Для этого они используют корпус страниц английской Википедии, автоматически размеченный по значениям ресурса BabelNet (включающего как составную часть WordNet) [16] – точность автоматической разметки оценивается как 77.8% [20]. Размеченный корпус используется следующим образом: неоднозначные слова получают дополнительную пометку суперпонятия, к которому они относятся. На основе полученного корпуса с разметкой суперпонятий авторы обучают векторные представления слов, слов с разрешенной неоднозначностью относительно суперпонятия и самих суперпонятий в одном векторном

пространстве word2vec [15]. В результате эмбединги для слов, обученных совместно с суперпонятиями, позволили добиться более высокого качества предсказания семантической близости слов [1, 12, 17] на четырех датасетах из пяти. Кроме того, эмбединги с суперпонятиями позволили улучшить результаты в задачах анализа тональности, классификации субъективных предложений и идентификации метафор.

В работе [2] исследуется порождение контекстуализированных эмбедингов для суперпонятий WordNet на основе нейросетевой модели BiLSTM. Авторы используют шесть обобщенных суперпонятий для представления существительных: Одушевленная сущность, Природный объект, Артефакт (созданный объект), Динамическая ситуация (например, действие), Статическая ситуация (например, состояние). Для каждого суперпонятия авторы отбирают по 200 соответствующих, а также случайным образом дополнительно отбираются негативные примеры, не относящиеся к данному суперпонятию. Полученные списки примеров используются для разметки корпуса, в котором слова из списков заменяются на позитивные или негативные теги каждого суперпонятия, получается псевдоаннотированный корпус.

На полученном корпусе обучается классификатор, который, получая слово и контекст, предсказывает вероятность отнесения слова к конкретному суперпонятию. Для классификации используется нейронная модель BiLSTM. В результате каждое слово получает некоторую величину в векторе размерности 6 суперпонятий. Таким образом, для каждого слова получается интерпретируемый вектор. Подход тестируется в задаче предсказания суперпонятия для слова на размеченном значениями французском корпусе и сравнивается с нейросетевой языковой моделью FlauBERT [11], французским вариантом модели BERT [6]. Показано, что на небольших объемах обучающих данных предложенный подход сравним с FlauBERT. Кроме того, метод намного быстрее и менее затратен по ресурсам, чем FlauBERT.

Как видно, в описанных работах суперпонятия (абстрактные семантические категории) заранее фиксированы. В нашей работе мы исследуем подход к гибкому (адаптивному) выделению суперпонятий для задачи или предметной области на основе соответствующего текстового корпуса и структуры соответствующих векторных представлений.

2. Идея адаптивных суперпонятий

Суперпонятия – это синсеты RuWordNet (или других ресурсов в форме семантических сетей), которые обладают следующими свойствами:

- суперпонятие должно иметь несколько гипонимов (видовых понятий), т.е. действительно являться обобщающим понятием для более конкретных понятий, представлять собой узел близких слов по тезаурусу,
- суперпонятие как узел близких по смыслу слов должно быть подтверждено и на основе некоторого представительного корпуса – а именно, слова из суперпонятий и связанных с ними понятий должны образовывать клику – граф, в котором все вершины связаны между собой. Связи между словами в клике считаются на основе косинусной меры близости эмбедингов по заданному корпусу, превышающей заданный порог. Например, синсет *город*, связанный в RuWordNet с большим количеством конкретных городов, соответствует клике из более, чем 200 названий-городов, чьи векторные представления близки друг к другу с косинусной мерой выше 0.4, посчитанной на основе корпуса Araneum.

Таким образом, суперпонятия – это семантические категории, выделенные людьми в семантических ресурсах, дополнительно подтвержденные высокой семантической близостью векторов, связанных с ними слов, что дает возможность адаптации набора суперпонятий на конкретную задачу или конкретную предметную область, т.е. предполагается, что для разных задач и связанных с ними корпусов данных набор суперпонятий должен различаться, отражая наиболее существенные категории в данной предметной области.

3. Алгоритм порождения суперпонятий

Алгоритм формирования суперпонятий включает следующие шаги:

1. Выделение из RuWordNet синсетов, которые имеют по крайней мере несколько видовых отношений (гипонимов) с другими синсетами – это кандидаты в суперпонятия;
2. Порождение множества слов (сематического хаба), соответствующих суперпонятию-кандидату – это все слова и выражения из синсета понятия в RuWordNet, а также слова и выражения из синсетов, непосредственно связанных с этим понятием отношениями гипоним-гипероним (родовидовыми отношениями);
3. Создание клики понятий для суперпонятия на основе косинусной близости эмбедингов синсетов хаба и заданного порога. Клики высчитываются с помощью алгоритма из пакета NetworkX. Вес связи между двумя синсетами определяется как средний вес связи слов из разных синсетов пары. Веса связей слова – это значение косинусной близости, посчитанной по некоторому корпусу дистрибутивной моделью (word2vec, glove). Усредненный вес должен превышать заданный порог. Чтобы был отобран как кандидат в суперпонятия, должна возникнуть минимальная клика – 3 понятия.
4. Для отобранного суперпонятия нужно создать представительное векторное представление, поскольку в семантическом хабе, связанном с суперпонятием (синонимы, гипонимы, гиперонимы) могут встречаться многозначные и/или редкие слова. Для этого из слов хаба создается сеть дистрибутивных отношений с весами косинусной близости, на сети вычисляется PageRank [5] каждого слова, отбираются пять наиболее весомых слова, и их усредненное векторное представление рассматривается как векторное представление суперпонятия. Примеры отобранных

суперпонятий на основе векторной модели корпуса Araneum [10] представлены в Таблице 1.

5. Следующим этапом является отбор суперпонятий из всех выделенных кандидатов. Отбор начинается с верхнего уровня иерархии семантической сети RuWordNet, поскольку мы заинтересованы в наиболее обобщающих, абстрактных понятиях. Если очередное суперпонятие похоже по векторной модели на уже отобранное и являющееся его гиперонимом, то это понятие отвергается, поскольку слишком много похожих суперпонятий приводят к запутанной системе семантических категорий.

Табл. 1. Примеры выделенных суперпонятий и их наиболее представительных слов

Идентификатор синсета RuWordNet	Название синсета RuWordNet	5 наиболее представительных слов, соответствующих суперпонятию
106509-N	Составная часть	<i>кусочек, полосочка, низ, полоска, бок.</i>
106501-N	Занятие, деятельность	<i>забава, занятие, развлечение, увлечение, отрасль</i>
106768-N	Свойство, характеристика	<i>абстрактность, безыскусность, стильность, отвлеченность, обособленность</i>
106646-N	Отношение между сущностями	<i>различие, соотношенность, взаимосвязь, несходство, несхожесть</i>

106505-N	Продукт труда	<i>подлинник, продукция, произведение, товар, издание</i>
----------	---------------	---

Упрощенный фрагмент лога отбора суперпонятий выглядит следующим образом:

```
{'106509-N',      'name':      'составная      часть',      'status':      'added'}
{'106646-N',      'name':      'отношение      между      сущностями',      'status':      'added'}
{'106846-N',      'name':      'вариант,      разновидность',      'status':      'added'}
{'132964-N',      'name':      'новшество',      'status':      'skipped',      'condition':      'clique_size < 3'}
..
{'112766-N', 'name': 'задняя часть', 'status': 'skipped', 'condition': 'close to (106509-N:0.59)'}
{'110735-N', 'name': 'частица (небольшая часть)', 'status': 'skipped', 'condition': 'close to (106509-N:0.50)'}
```

Из лога видно, что часть суперпонятий не одобряется (skipped) из-за отсутствия соответствующей клики ('clique_size < 3'), а часть суперпонятий отвергается из-за слишком высокого сходства с уже отобранными суперпонятиями ('close to').

Несмотря на сделанный отбор суперпонятий, выделяются еще близкие по смыслу суперпонятия, которые обычно представляют собой подвиды более широкого понятия (понятия-сестры), например, понятия *гуманитарные науки* и *общественные науки*. В таком случае естественно обобщать понятия-сестры к более широкому понятию. Приведем примеры обобщений похожих по векторной модели понятий-сестер к родовому понятию:

- *гуманитарные науки, общественные науки → наука*
- *научное учреждение, высшее учебное заведение → организация, выполняющая исследования и разработки*

- *иностранное государство, европейская страна, развитое государство, республика (государство), социалистическое государство → государство*
- *цветковое растение, травянистое растение, кустарник, вечнозеленое растение → растение*
- *воинское звание, дворянский титул → звание*

В зависимости от корпуса, на котором рассчитывается векторная модель, описанная процедура дает 100-300 суперпонятий, объединяющих разное количество элементов. Суперпонятий (семантических классов) не должно быть много. Поэтому на последнем шаге суперпонятия пополняются всеми остальными словами из корпуса (на текущем этапе только существительными из RuWordNet, присутствующими в корпусе), косинусное сходство которых с суперпонятиями выше некоторого порога, и отбирается заданное количество суперпонятий с максимальным количеством элементов.

Таким образом, в результате описанной процедуры отобраны суперпонятия (семантические категории) из ресурса RuWordNet, которым одновременно соответствует некоторое множество слов, близких между собой по корпусной дистрибутивной модели (клика). Каждое суперпонятие имеет свое векторное представление по исходной дистрибутивной модели. Все остальные слова из текстового корпуса можно представить в виде вектора схожести на эти суперпонятия, т.е. если отобрано 100 суперпонятий, то для каждого слова будет получен вектор размерности 100 интерпретируемых измерений.

4. Эксперименты

В эксперименте по описанному алгоритму использовался тезаурус RuWordNet и векторные модели, обученные на корпусе из 800 тысяч новостных статей. На основе близости слов к векторам суперпонятий каждому слову из векторной модели был сопоставлен вектор из K компонент по векторным близостям слова к суперпонятиям.

В данном эксперименте применялись следующие параметры:

1. Параметры алгоритма построения списка суперпонятий:

- a. Минимальное количество родовидовых отношений у понятия для формирования суперпонятия (=5),
- b. Минимальное значение близости между понятиями для формирования ребра в процессе построения клики суперпонятия (=0.4),
- c. Минимальное значение близости между словами для формирования ребра между словами в суперпонятии для построения графа и последующего расчета pagerank на нем (=0.4),
- d. Минимальный размер клики из понятий в суперпонятии (=3),
- e. Минимальный размер суперпонятия в словах (=10),
- f. Количество слов для расчета вектора суперпонятия на основе PageRank (=5),
- g. Минимальное значение близости между добавляемым суперпонятием и его уже добавленными гиперонимами при котором суперпонятие будет добавлено в итоговый список (=0.3),

2. Параметры алгоритма обобщения:

- a. Минимальное значение близости между парой ко-гипонимов для обобщения на гипероним (=0.4),
- b. Минимальная глубина суперпонятия, которое может служить обобщением (=3),
- c. Количество слов для расчета вектора суперпонятия (=5).

Приведем примеры полученных суперпонятий RuWordNet на данных параметрах: *составная часть; продукт труда; занятие, деятельность; пребывать в состоянии; изображение*

(результат); группа, объединенная по общему признаку; субъект деятельности; вещество; научный работник и др (см. Приложение). Как видно, получилось выделить обобщающие семантические категории. Также можно заметить, что выделяются и более узкие категории, например, среди понятия *субъект деятельности* отдельно выделяется категория *научный работник*, которая имеет специфические контексты. В этом и состоит предполагаемая адаптивность метода выделения суперпонятий, который может выделить и существенные, важные, специфичные категории не обязательно самого верхнего уровня семантической сети.

В качестве эксперимента проверялась возможность отображать разные модели в одно интерпретируемое векторное пространство. Для этого на основе вышеупомянутого корпуса новостей были обучены две векторные модели word2vec с одинаковыми параметрами (skip-gram, размер окна 3, количество эпох 10).

Первой частью эксперимента была проверка, что две векторные модели, обученные даже на идентичном корпусе и с одинаковыми параметрами, будут иметь разные векторные представления. Для этого, для каждого слова из коллекции считалась косинусная близость между векторами из двух разных моделей. Гистограмма близостей слов представлена на Рис. 1. Средняя близость оказалась равна 0.6. Это означает, что обычные векторные модели, даже обученные на одном корпусе, создают несопоставимые векторные представления, в которых представления даже одного и того же слова могут быть не похожи друг на друга.

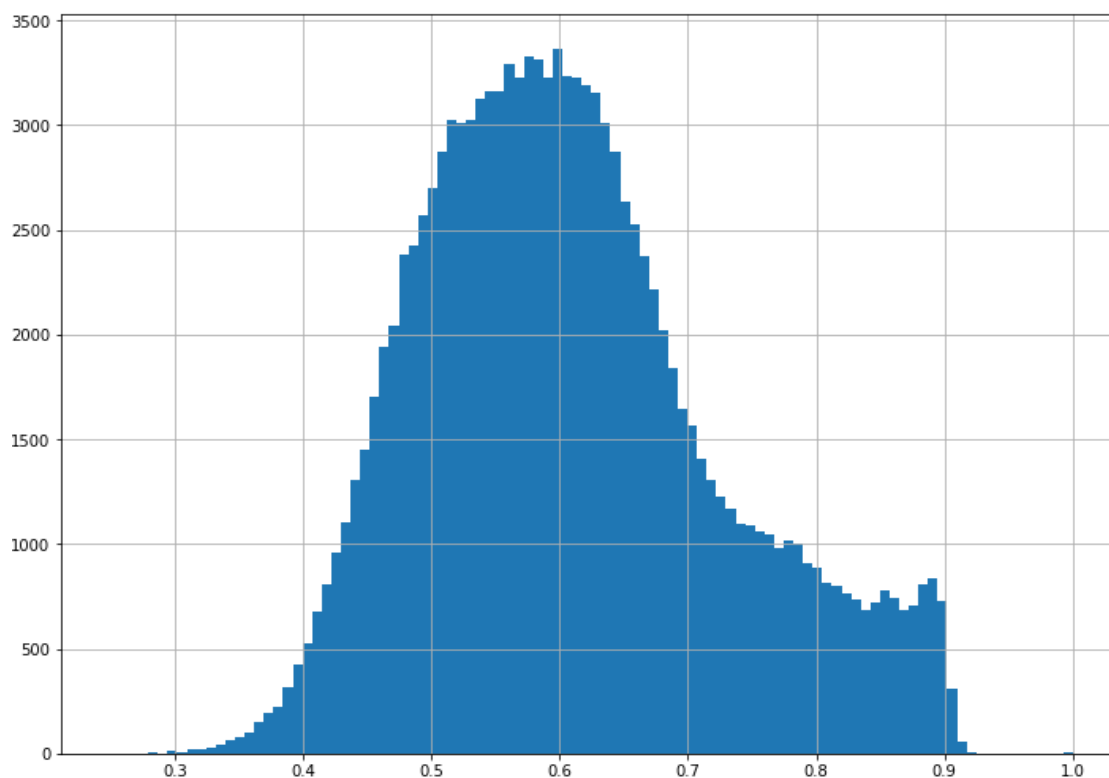


Рисунок 1: Гистограмма близостей слов к самим себе в разных моделях. По x – мера близости, по y – количество слов

Для проверки сохранения семантической близости на основе векторов суперпонятий мы исследовали, насколько такое представление может выявлять сходство между семантически близкими словами из независимо обученных векторных моделей. Для этого предварительно был составлен список из ~100 пар слов, на основе ранее обученной модели word2vec в той же постановке, которые имели высокое сходство. С этой целью для каждого слова из коллекции был получен список наиболее близких слов, эти списки были упорядочены по наивысшему сходству среди полученных близких слов. Для первых 100 слов в списке были выбраны их наиболее близкие слова.

Используя полученный список, при обучении векторных моделей, в первом случае из корпуса удалялись все «правые» слова пар, во втором случае все «левые» слова пар. Таким образом, итоговые словари векторных моделей отличались в соответствии с полученным списком пар. Например, для пары (**беларусь**, **белоруссия**) при обучении первой векторной модели, были удалены все случаи использования слова **белоруссия**, а при обучении второй модели, все случаи использования **беларусь**. Таким образом, эти модели не имеют в своих словарях соответствующих слов.

Далее, для каждой из двух векторных моделей, в соответствии с изложенным алгоритмом построения, были получены векторные представления на основе суперпонятий. Для того, чтобы полученные векторные модели были полностью сопоставимы, на нескольких этапах алгоритма было выполнено выравнивание:

1. Вес максимальной клики для каждого суперпонятия был получен усреднением соответствующих весов в двух моделях,
2. Итоговый список суперпонятий формировался как пересечение списков суперпонятий, полученных в двух моделях, а позиция суперпонятий формировалась как средняя позиция.

Таким образом, полученные списки суперпонятий получились идентичными и содержали в себе 78 суперпонятий. Так как итоговое количество суперпонятий из-за процедуры пересечения списков получилось меньше 100, то операция по усечению множества не проводилась. В более простом случае, когда работа идет только с одной моделью, обычно количество суперпонятий получается более 100.

В итоге каждое слово из модели было отображено в векторную модель на основе суперпонятий, и, если слово встречалось в обеих моделях, его представление усреднялось. В полученной векторной модели словарь является объединением словарей двух моделей.

Для проверки гипотезы, что с помощью разработанного подхода возможно отображать разные модели в одно векторное пространство был проведен анализ того, насколько близко к друг другу располагаются слова из выделенных ранее пар, по отношению к другим словам.

Для этого, в полученной векторной модели для каждого слова из пары производилось вычисление списка наиболее близких слов, среди которых происходил поиск второго слова из пары и его позиции в списке. Например, 5 наиболее близких слов к слову **беларусь**: **белоруссия** 0.991, казахстан 0.988, судан 0.984, украина 0.984, россия 0.984, а к слову белоруссия: беларусь 0.991, казахстан 0.984, латвия 0.983, венгрия 0.983, молдавия 0.983.

Таким образом, средний ранг для пары (**беларусь, белоруссия**) равен 0 (отсчет идет с 0).

Результаты показали, что для 60% пар, искомое слово оказывалось в топ-5. Итоговое распределение позиций можно видеть на Рис. 2.

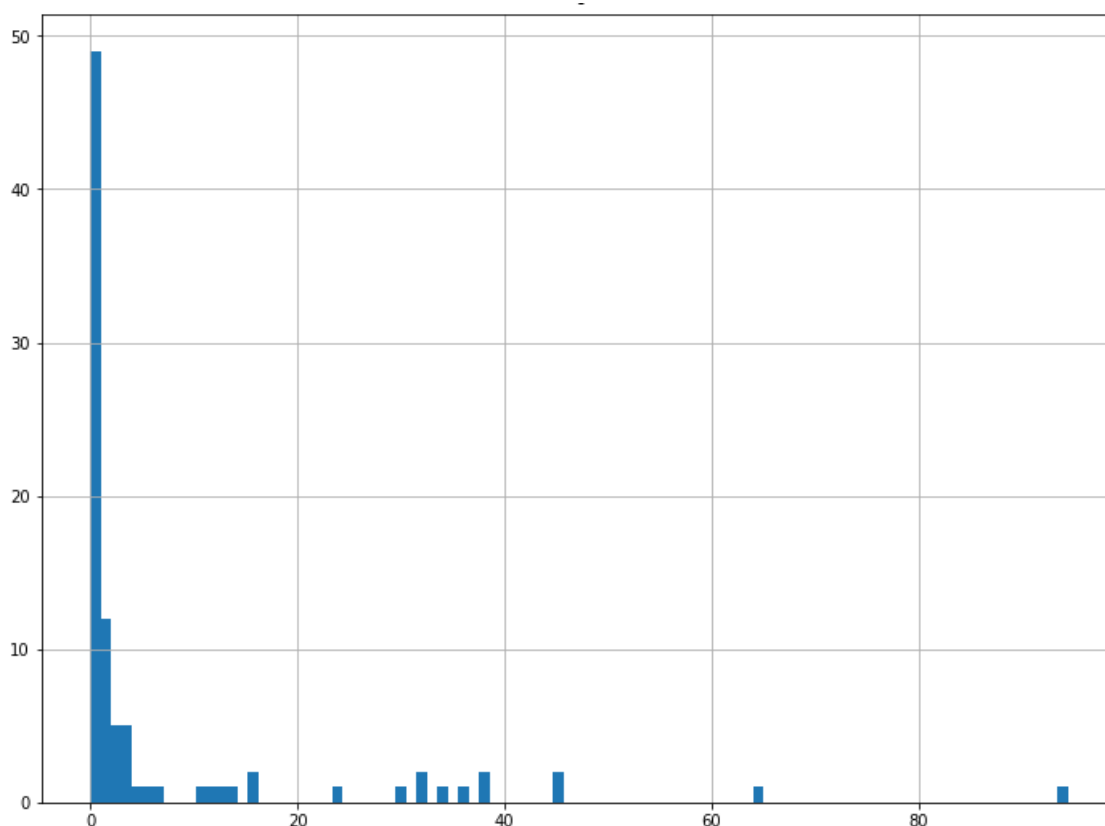


Рисунок 2: Гистограмма распределения рангов для исследуемых пар слов. По x – ранг, по y – количество пар

Таким образом, на корпусе были независимо обучены две модели, в каждой из которых не содержалось некоторое множество слов, близких друг к другу, но в представлении на основе суперпонятий эти слова оказались близки.

Заключение

В данной работе представлен подход по созданию интерпретируемых векторных представлений слов, в которых каждая компонента вектора соответствует некоторой интерпретируемой семантической категории. Для получения таких категорий используется лексико-семантический ресурс в виде семантической сети русского языка RuWordNet, а также представительный корпус русскоязычных текстов для порождения векторных представлений.

Полученные интерпретируемые векторные представления были протестированы на наличие возможности отображать разные модели в одно векторное пространство таким образом, чтобы близкие по смыслу слова, которые присутствуют в одной модели, но отсутствуют в другой, оказывались близко друг к другу в итоговой векторной модели.

Финансирование

Работа поддержана грантом исследовательских центров по искусственному интеллекту (соглашение 000000D730321P5Q0002 No. 70-2021-00142 с ИСП РАН).

Конфликт интересов

Процесс написания и содержание статьи не дает оснований для постановки вопроса о конфликте интересов.

Соответствие этическим стандартам

Данная статья является полностью оригинальным произведением ее авторов, прежде не опубликована и до получения решения редколлегии PRIA о непринятии ее к публикации не будет направляться в другие издания

Список литературы

1. E. Agirre, E. Alfonseca, K. Hall, J. Kravalová, M. Pasca, and A. Soroa, “A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches”, Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. (2009).
2. C. Aloui, C. Ramisch, A. Nasr, and L. Barque, “Slice: Supersense-based lightweight interpretable contextual embeddings”, In Proceedings of the 28th International Conference on Computational Linguistic. (2020).

3. Ю.Д. Апресян, “Лексическая семантика: Синонимические средства языка”, Москва, (1974).
4. M. Artetxe, G. Labaka and E. Agirre, “Learning principled bilingual mappings of word embeddings while preserving monolingual invariance”, Proceedings of the 2016 conference on empirical methods in natural language processing. (2016).
5. S. Brin and, L. Page, “The anatomy of a large-scale hypertextual web search engine”, Computer networks and ISDN systems. Т. 30, №. 1-7 (1998).
6. J. Devlin, W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). (2019).
7. C. Fellbaum, “WordNet: An Electronic Lexical Database”, Cambridge, MA., MIT Press (1998).
8. L. Flekova, I. Gurevych, “Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization”, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2016).
9. И. М. Кобозева, “Лингвистическая семантика”, Москва, УРСС, (2004).
10. A. Kutuzov, M. Kunilovskaya, “Size vs. structure in training corpora for word embedding models: Araneum rassicum maximum and russian national corpus”, International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, (2017).
11. H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, ... and D. Schwab, “FlauBERT: Unsupervised Language Model Pre-training for French,” In Proceedings of the 12th Language Resources and Evaluation Conference. (2020).
12. I. Leviant, R. Reichart, “Separated by an un-common language: Towards judgment language informed vector space modeling”, arXiv preprint arXiv:1508.00106. (2015).

13. N.V. Loukachevitch, G. Lashevich, A.A. Gerasimova, V.V. Ivanov, and B.V. Dobrov, “Creating Russian wordnet by conversion”, Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”. (2016).
14. T. Mikolov, Q.V. Le, I. Sutskever, “Exploiting similarities among languages for machine translation”, arXiv preprint arXiv:1309.4168. (2013).
15. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality”, In Advances in Neural Information Processing Systems. T. 26 (2013).
16. R. Navigli, S. P. Ponzetto, “BabelNet: Building a very large multilingual semantic network”, Proceedings of the 48th annual meeting of the association for computational linguistics. (2010).
17. A. Panchenko, D. Ustalov, N. Arefyev, D. Paperno, N. Konstantinova, N. Loukachevitch, and C. Biemann, “Human and machine judgements for Russian semantic relatedness”, In International conference on analysis of images, social networks and texts. (2016).
18. M. Radovanovic, A. Nanopoulos, M. Ivanovic, “Hubs in space: Popular nearest ‘ neighbors in high-dimensional data”, The Journal of Machine Learning Research. (2010).
19. S. Ruder, I. Vulić, A. Søgaard, “A survey of cross-lingual word embedding models”, Journal of Artificial Intelligence Research. T. 65 (2019).
20. F. Scozzafava, A. Raganato, A. Moro, & R. Navigli, “Automatic identification and disambiguation of concepts and named entities in the multilingual wikipedia”, In Congress of the Italian Association for Artificial Intelligence. (2015).
21. J. Yamane, T. Takatani, H. Yamada et al., “Distributional hypernym generation by jointly learning clusters and projections”, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. (2016).

В приложении приводится список суперпонятий для модели, которая была получена в описанном эксперименте:

- Суперпонятие 0: составная часть: полосочка, кусочек, мякоть, бочок, компонент
- Суперпонятие 1: продукт труда: произведение, продукция, товар, продукт, изделие
- Суперпонятие 2: занятие, деятельность: занятие, отрасль, хобби, профессия, увлечение
- Суперпонятие 3: пребывать в состоянии: томление, жизнелюбие, преклонение, оттаивание, голодание
- Суперпонятие 4: изображение (результат): фотография, рисунок, изображение, гравюра, снимка
- Суперпонятие 5: группа, объединенная по общему признаку: тройка, тридцатка, десяток, сотня, двадцатка
- Суперпонятие 6: субъект деятельности: должник, оппонент, покупатель, благотворитель, организатор
- Суперпонятие 7: вещество: отдушка, консервант, порошок, ингредиент, вещество
- Суперпонятие 8: измениться, изменение: улучшение, увеличение, ослабление, замедление, пропитка
- Суперпонятие 9: место в пространстве: сгиб, выпуклость, пятнышко, поверхность, стремнина
- Суперпонятие 10: природное явление: погода, осадка, облако, ветер, циклон
- Суперпонятие 11: истратить, израсходовать: расход, экономия, потребление, перерасход, затрата
- Суперпонятие 12: компьютерная программа: утилита, приложение, подпрограмма, программа, апдейт
- Суперпонятие 13: биологическая сущность: клетка, ткань, митохондрия, организм, хромосома
- Суперпонятие 14: движение, перемещение: катание, передвижение, наклон, путешествие, скачок
- Суперпонятие 15: состояние, внутренние обстоятельства: упорядоченность, сухость, влага, мокрота, ситуация
- Суперпонятие 16: население: горожанин, латиноамериканка, европейка, уралец, свердловчанин
- Суперпонятие 17: физиологический процесс: климакс, регенерация, секреция, сердцебиение, менопауза
- Суперпонятие 18: физическое свойство: проницаемость, густота, проводимость, ломкость, свойство
- Суперпонятие 19: единица измерения информации: мбайт, гигабайт, мбитый, кбит, мегабайт
- Суперпонятие 20: единица объема: галлон, декалитр, поллитра, баррель, литр
- Суперпонятие 21: изменить, сделать иным: улучшение, осветление, увеличение, уменьшение, изменение
- Суперпонятие 22: единица массы: килограмм, грамм, кило, тонна, унция
- Суперпонятие 23: способность: тактичность, музыкальность, умение, гибкость, расположенность
- Суперпонятие 24: материал для изготовления: поролон, латекс, пластик, обмотка, резина
- Суперпонятие 25: единица длины: микрометр, метр, сантиметр, нанометр, миллиметр
- Суперпонятие 26: сооружение, постройка: сооружение, здание, постройка, шалаш, строение
- Суперпонятие 27: предмет, вещь: мотка, брусочек, вещица, шутовина, украшение
- Суперпонятие 28: возраст: отрочество, детство, возраст, свойство, старость
- Суперпонятие 29: форма, наружный вид: округлость, выпуклость, комплекция, покрой, фасон

Суперпонятие 30: бог: бог, господь, христос, богочеловек, иисус

Суперпонятие 31: метод лечения: гемодиализ, терапия, диализ, рефлексотерапия, анестезия

Суперпонятие 32: высказывание (то, что высказано): высказывание, рассуждение, реплика, шуточка, изречение

Суперпонятие 33: чувство, эмоция: чувство, переживание, гордыня, разочарование, радость

Суперпонятие 34: черта характера: сердечность, интеллигентность, уступчивость, уравновешенность, непоколебимость

Суперпонятие 35: географическая область: приангарий, зауралье, урал, приамурья, прибайкалье

Суперпонятие 36: создать (сделать существующим): протирание, кройка, отливка, ткачество, прокалывание

Суперпонятие 37: небесное тело: комета, астероид, планета, солнце, луна

Суперпонятие 38: административно-территориальная единица: анлия, шотландия, волость, уезд, уэльс

Суперпонятие 39: произнести, выговорить, проговорить: разглагольствование, причитание, произнесение, трескотня, акцентирование

Суперпонятие 40: мнение: трагичность, трагизм, предположение, суждение, истинность

Суперпонятие 41: исследование, изучение: изучение, исследование, анализ, изыскание, слежение

Суперпонятие 42: промежуток времени: месяц, час, полгода, век, полчаса

Суперпонятие 43: вооруженное формирование: спецназ, отряд, подразделение, войско, спецотряд

Суперпонятие 44: важный, значительный: важность, всемогущество, могущество, величие, приоритетность

Суперпонятие 45: положение, тезис: догмат, постулат, суждение, мысль, принцип

Суперпонятие 46: переместить: опускание, переброс, тряска, размахивание, сваливание

Суперпонятие 47: государство: индия, джибути, египет, кения, малайзия

Суперпонятие 48: созвездие зодиака: водолей, овен, телец, скорпион, стрелец

Суперпонятие 49: высшее учебное заведение: вуз, сельхозинститут, университет, госуниверситет, факультет

Суперпонятие 50: воинское звание: лейтенант, сержант, подполковник, майор, ефрейтор

Суперпонятие 51: стиль в искусстве: модернизм, романтизм, минимализм, реализм, классицизм

Суперпонятие 52: предприятие общественного питания: ресторан, кафе, ресторанчик, бистро, кафешка

Суперпонятие 53: массовая акция: шествие, митинг, пикет, акция, марш

Суперпонятие 54: физические упражнения: упражнение, трусца, подтягивание, гимнастика, разминка

Суперпонятие 55: скульптурное произведение: изваяние, монумент, скульптура, памятник, статуя

Суперпонятие 56: полуостров: камчатка, таймыр, полуостров, аляска, флорида

Суперпонятие 57: изобразить (воспроизвести): вышивание, резьба, тиснение, гравировка, вышивка

Суперпонятие 58: издать звук: урчание, завывание, шипение, дрожание, шуршание

Суперпонятие 59: достижения в деятельности: успех, достижение, вклад, итог, лепта

Суперпонятие 60: кормить, давать еду: корм, прикорм, подкормка, откорм, прикормка

Суперпонятие 61: план деятельности: программа, госпрограмма, подпрограмма, перечень, список

Суперпонятие 62: марка автомобиля: газель, мазд, ситройня, жигули, фольксваген

Суперпонятие 63: приготовление пищи: выпекание, запекание, выпечка, копчение, варка

Суперпонятие 64: военнослужащий: военнослужащий, танкист, кавалерист, офицер, десантник

Суперпонятие 65: водные средства транспорта: судно, корабль, теплоход, катамаран, буксир

Суперпонятие 66: экономический показатель: рентабельность, пассажирооборот, грузооборот, себестоимость, показатель

Суперпонятие 67: деньги: деньга, сумма, расход, оплата, задолженность

Суперпонятие 68: драка: драка, потасовка, стычка, конфликт, поножовщина

Суперпонятие 69: перелетная птица: куропатка, чирок, скворец, аист, перепел

Суперпонятие 70: город иркутской области: слюдянка, шелехов, иркутск, нижеудинск, черемхово

Суперпонятие 71: научный работник: естествоиспытатель, историк, профессор, этнолог, филолог

Суперпонятие 72: курортный город: кисловодск, пятигорск, сочи, евпатория, железноводск

Суперпонятие 73: охотничья собака: собака, собачонка, собачка, спаниель, терьер

Суперпонятие 74: заместитель должностного лица: заместитель, замглавы, зам, замминистра, замначальник

Суперпонятие 75: награда: вознаграждение, награда, медаль, приз, премия

Суперпонятие 76: записать на специальный носитель: фиксация, видеозапись, запись, аудиозапись, видеофиксация

Суперпонятие 77: рыба: толстолобик, камбала, рыба, скумбрия, корюшка