

УДК 004.89

DOI 10.17726/phillT.2024.1.5



**Новый способ нахождения аналогов
как возможность исследования языка,
мышления и построения систем
искусственного интеллекта**

Хомяков Александр Борисович,

*независимый исследователь,
магистр физических наук
Санкт-Петербург, Россия
alexander.xom@gmail.com*

Чижик Петр,

*независимый исследователь,
магистр информатики
Германия*

Аннотация. В статье представлен новый способ получения аналогов слов, отличающийся простотой и отсутствием необходимости предварительного обучения на больших данных, как в существующих методах. В предложенном методе аналоги определяются по их синтаксическим предикатам с применением методов дистрибутивной семантики. В исследовании были получены и проанализированы аналоги прилагательных, существительных и глаголов. Это позволило прийти к результату, который не уступает результатам, полученным с помощью наиболее популярного нейросетевого подхода word2vec, при качественном сравнении аналогов. Демонстрируемый метод показывает, что получение аналогов возможно на методах дистрибутивной семантики с применением более интерпретируемого метода, что открывает возможность исследования семантической аналогии. Данный метод позволяет также определять аналоги по определенной тематике. На основе полученных экспериментальных результатов формулируется оригинальное определение аналогов и когнитивных схем. Также в статье анализируется и обосновывается возможность нового подхода для создания систем искусственного интеллекта на основе исследованного метода. По мнению авторов, это дает значительные преимущества для создания таких систем. В частности, предлагаемый метод позволяет создавать более широкие обобщения на порядки меньших данных, а также обучение во время использования, что недоступно для нейросетей.

Ключевые слова: дистрибутивная семантика; нейросети; компьютерная лингвистика; word2vec; аналоги; синонимы; семантическая близость; искусственный интеллект.

A new way of finding analogues as an opportunity to study language, thinking and build artificial intelligence systems

Khomyakov Alexander Borisovich,

*independent researcher,
Master of Physical Sciences
Saint-Petersburg, Russia*

alexander.xom@gmail.com

Chizhik Petr,

*independent researcher,
Master of Computer Science
Germany*

Abstract. The article presents a new method for obtaining analogues of words, characterized by simplicity and the absence of the need for preliminary training on large data as in existing methods. In the method under study, analogues are determined by their syntactic predicates using methods of distributive semantics. In the study, analogues of adjectives, nouns and verbs were obtained and analyzed. This made it possible to obtain a result that is not inferior to the results obtained using the most popular neural network approach as word2vec when qualitatively comparing analogues. The demonstrated method shows that obtaining analogues is possible using methods of distributive semantics using a more interpretable method, which opens up the possibility of studying semantic analogy. This method also allows you to identify analogues on a specific topic. Based on the experimental results obtained, an original definition of analogues and cognitive schemes is formulated. The article also analyzes and substantiates the possibility of a new approach for creating artificial intelligence systems based on the researched method. According to the authors, this provides significant advantages for the creation of such systems. In particular, the proposed method allows for broader generalizations over orders of magnitude smaller data, as well as learning during use, which is not possible for neural networks.

Keywords: distributional semantics; neural networks; computational linguistics; word2vec; analogues; semantic proximity; AI.

Цель работы

Способы получения аналогов существуют с 2013 года, со времени появления word2vec [1], который относится к нейросетевым подходам. Недостатками такого подхода являются неполнота, неконтролируемость подборок аналогов и необходимость больших данных для обучения, невозможность снятия омонимии в аналогах. Чтобы устранить эти недостатки в получении аналогов, мы применили классическое определение аналогов как схожих по признакам, определив признак как синтаксический предикат. Красивая (adj) яблоня (noun), яблоня (subj) расцветает (verb), расцветает (verb) в саду (obj) – это предикаты. Аналог мы определяем как слово с наибольшим числом схожих предикатов. Целью являлась проверка возможности получения аналогов по признакам, в качестве которых были взяты синтаксические предикаты, что позволяет применять к ним методы дистрибутивной лингвистики. Это дает возможность исследовать аналогию на интерпретируемых результатах – по общим предикатам.

На сегодня существует проблема в определении аналогов, для которых пока имеется лишь качественная оценка того, что считать аналогом или синонимом слова. Синонимы определяются только словарями синонимов, составленными лингвистами. Проверяемой и измеряемой меры аналогии нет, хотя были исследования способа получения аналогов в word2vec [3-5], которые сводятся к тому, что слова употребляются в похожем контексте. Учитывая, что аналогия является одним из важнейших элементов в моделировании мышления [11], авторам видится важность открытия механизмов ее формирования.

Материалы и методы исследования

Для исследования были выбраны тексты книг (подборка книг художественной литературы на русском языке, 200 книг). Далее были выделены из текстов парсером грамматики зависимостей LinkParser предикаты в виде пар подлежащего и сказуемого, сказуемого и дополнения, а также именные группы. Все слова лемматизировались. Таким образом были получены 1 442 924 пары «подлежащее-сказуемое», 136 500 пар «сказуемое-дополнение» и 821 000 пар «прилагательное-существительное», для каждой пары указывалась частота встречаемости при парсинге текста.

Из этих пар были созданы вектора предикатов (subject; predicate) каждого слова вида $\text{subj1}\{\text{verb 1, verb 2, ... verb n}\}$, $\text{subj2}\{\text{verb 1, verb 3, ... verb n}\}$, $\text{subj3}\{\text{verb 2, verb 5, ... verb n}\}$ на едином словаре лемматизированных слов. В методике исследования выделение аналогов определялось по наибольшему числу общих предикатов. Для определения были взяты не все, а только первые 50 частотных предикатов. Таким методом было выделено по 5000 лексем с аналогами для прилагательных, существительных и глаголов.

В результате исследования было замечено, что существительные отличаются именно теми предикатами, которые идут по частоте встречаемости в текстах после наиболее распространенных (например, распространенные глаголы «стать», «иметь», «быть», «являться»). Именно менее распространенные предикаты являются дифференцирующими для поиска аналогичных лексем. Поэтому для определения аналогов по таким дифференцирующим их предикатам наиболее подходящим из всех опробованных методов оказался известный метод TF-IDF (от англ. TF – term frequency, IDF – inverse document frequency). В нем рассматривается обратная величина от встречаемости слов в документах, умноженная на частоту слова в документе. Таким образом, чем чаще слово встречается во всех документах, тем ниже важность слова для данного документа, то есть она меньше его характеризует, чем реже встречающееся в других документах. Документом в нашем исследовании выступает вектор предиката $\text{subj1}\{\text{verb 1, verb 2, ... verb n}\}$, а словами – слова из векторов. Так как у нас все слова в векторах предикатов встречаются только раз, применялась только одна часть метода – IDF без TF как меры частоты слова в одном документе. Поэтому мы назвали свой метод PredicatIDF, то есть метод IDF, построенный на векторах предикатов.

Результаты исследования

Результаты исследования показали, что таким способом даже на ограниченных данных можно получить достаточно точные аналоги. В качестве метрики было взято качественное сравнение с синонимами, полученными методом word2vec в проекте Rusvectors.org на словаре НКРЯ (Национальный корпус русского языка). При качественном сравнении аналогов методом PredicatIDF критерием было наличие таких же лексем среди первых 10 лексем

по word2vec в Rusvectors.org. Чаще всего аналоги по методу PredicatIDF были разнообразнее, как показано ниже. Вот примеры сравнения:

PredicatIDF

Время: час: 3.26, минута: 2.77, день: 2.45, зима: 2.34, лето: 2.20, сутки: 2.02, ночь: 2.00, секунда: 1.97, вечер: 1.96, век: 1.95, мгновение: 1.89, период: 1.88, осень: 1.88, неделя: 1.87, месяц: 1.84, пора: 1.82, год: 1.81, срок: 1.74, конец: 1.74, эпоха: 1.74, миг: 1.68, утро: 1.64, весна: 1.62, очередь: 1.58, май: 1.57, момент: 1.55, трамвай: 1.53, праздник: 1.48, полоса: 1.45, деньга: 1.45

[Rusvectors.org/ru/associates/#](https://rusvectors.org/ru/associates/#)

Время: времени 0.56, пора 0.55, период 0.54, десятилетие 0.47, момент 0.47, эпоха 0.45, промежуток 0.39, срок 0.37, обстоятельство 0.35, оледенение 0.35

PredicatIDF

Проблема: вопрос: 3.18, противоречие: 2.91, конфликт: 2.76, ситуация: 2.32, загадка: 2.22, трудность: 2.21, задача: 2.16, проблематика: 1.96, разногласие: 1.94, возможность: 1.89

[Rusvectors.org/ru/associates/#](https://rusvectors.org/ru/associates/#)

Проблема: вопрос 0.71, задача 0.68, проблематика 0.65, аспект 0.61, тема 0.56, ситуация 0.55, конфликт 0.55, подход 0.52, трудность 0.51, частность 0.50

В результате качественной оценки были получены следующие результаты:

Аналоги существительных	Точность 87%
Аналоги глаголов	Точность 84%
Аналоги прилагательных	Точность 95%

Наилучшие результаты получены по аналогам прилагательных (по предикатам с вершинами именных групп). Вот пример получаемых аналогов методом PredicatIDF:

Голубой: синий: 3.51, желтый: 2.91, алый: 2.81, фиолетовый: 2.59, зеленый: 2.58, белый: 2.48, кроваво-красный: 2.43, черный: 2.38, многоцветный: 2.35, пурпурный: 2.34, ослепительный: 2.30, голубоватый: 2.30, красный: 2.29, оранжевый: 2.24, серебристый: 2.16, зеленоватый: 2.15, красноватый: 2.15, розоватый: 2.10, сверкающий: 2.08, багровый: 2.04, разноцветный: 2.03, слепящий: 2.02, смоляной: 1.99, темно-синий: 1.99, багряный: 1.97, яркий: 1.95, розовый: 1.92, темно-красный: 1.92, золотистый: 1.90, серый: 1.84, янтарный: 1.81, сизый: 1.81, желтоватый: 1.80.

Железный: металлический: 3.22, чугунный: 2.78, стальной: 2.76, кованый: 2.64, деревянный: 2.26, массивный: 1.98, железобетонный: 1.95.

Сравнение аналогов прилагательных показало различие результатов PredicatIDF и word2vec так как были более разнообразные. Так же интересные и абсолютно точные результаты были получены на числительных.

Первый: второй: 3.78, третий: 3.63, четвертый: 3.12, последний: 3.04, следующий: 2.88, шестой: 2.75, седьмой: 2.60, восьмой: 2.46, очередной: 2.44, пятый: 2.43, девятый: 2.42, двенадцатый: 2.27, одиннадцатый: 2.20, семнадцатый: 2.50, считанный: 02.40, десятый: 1.98, каждый: 1.96, девятнадцатый: 1.96, предпоследний: 1.89, последующий: 1.86.

Первичные данные и код проекта находятся по адресу <https://github.com/smer44/metaphors>.

Обсуждение

Надо отметить, что аналоги не являются в полном смысле синонимами. Таким методом нами получены именно аналоги по возможности применения в тех же контекстах, что и исходное слово с дифференцирующими аналоги предикатами. В них поэтому попадают и антонимы.

Лексемы существительных при определении аналогов по синтаксическим предикатам были взяты в роли как подлежащего, так и дополнения глагола. Сравнение результатов показало, что аналоги существительных лучше определяются в парах «сказуемое и дополнение», то есть как возможные действия над объектом (существительным). Разница получилась значительной: 87% точности аналогов в парах «сказуемое и дополнение» против 67% в парах «подлежащее и сказуемое» – что является предметом для анализа. Также были опробованы смешанные способы, когда в предикаты добавляются предикаты с определениями (прилагательные) или предикаты с подлежащим и дополнением вместе. Это не позволяло добиться каких-либо улучшений.

Представляемый метод дает возможность искать аналоги при помощи программ поиска по текстам, как ElasticSearch, например, для поиска аналогов по определенной тематике, отбирая аналоги по конкретным предикатам, исключая предикаты другого значения.

Например, слово «предмет» имеет аналоги как физический объект (значение 1, аналоги – объект, вещь), а также как предмет знаний (значение 2, аналоги – научная тема, область знаний). И в этих разных значениях они имеют разные общие глаголы с аналогами. В значении 1 это будут глаголы «положить, переместить, сдвинуть», а в значении 2 – «преподавать, изучать, докладывать».

Метод требует дальнейших исследований и уточнения получаемых аналогов за счет более точных парсеров, так как зависит от их качества, а также от устойчивых выражений, которые создают неодинаковое распределение частотности в выборке первых 50 предикатов для анализа. Для этого можно использовать для получения аналогов предложенным способом в несколько раз большие тексты (2000-3000 книг), чтобы получить более выраженное распределение частотности. Также есть возможность уточнения за счет исключения «мусорных» предикатов – с именами собственными, фразеологических оборотов, идиом. Но в целом представленное исследование показывает, что аналоги можно получать иным, более простым способом с приемлемым качеством, с возможностью управления подбором аналогов по теме или текущему тексту при быстром извлечении с помощью парсера дополнительных предикатов.

Значение результатов для исследования языка и мышления

Предложенный метод можно использовать для определения семантической близости так же, как в word2vec, – в виде коэффициента близости любого слова к заданному. Но при этом слова должны иметь хотя бы один общий предикат. Таким обычно являются распространенные глаголы «становиться», «быть», «являться», «находиться» и т.п. Сходство только по ним будет самым маленьким, так как по IDF эти глаголы есть в большинстве «документов» (векторах предикатов), то есть небольшое сходство есть у всех слов по данной методике.

Нахождение семантической близости с обязательным присутствием или отсутствием заданных предикатов является новым исследовательским приемом для широких исследований, которые более интерпретируемы, чем существующие. Например, мы исследовали все предикаты с одним распространенным глаголом «на-

ходиться» в одном тематическом домене – дом и все, что в нем. В результате мы получили карту всего, что находится в доме: комнаты, кухня, спальня, стол, диван, холодильник и т.п. Карта дома складывается по общему распространенному предикату и семантической близости (семантической группировке) слов. Более того, с помощью последовательного подбора актантов предиката «находиться» можно построить переходы из любого места в доме в любое другое. Это очень интересный эффект, который может позволить создавать на лингвистической статистике когнитивные карты как предполагаемый способ работы мышления.

Интересно, что до настоящего момента нет определения лингвистических аналогов. Общее определение как «схожих по признакам» не всегда применимо, так как мы имеем дело со словами, а не с обозначаемыми ими сущностями. В представленном методе сходство найдено не по признакам самих обозначаемых словом предметов, а по схожести предикатов, с которыми они синтаксически связаны. Существительные – по глаголам, которые действуют над ними, прилагательные – по существительным, к которым они относятся. И эти предикаты всегда множественны. Мы только интуитивно ощущаем, что два слова являются аналогами, так как не можем представить одновременно сразу все предикаты, которые создают аналогию.

Данное исследование открывает дискуссию о том, что такое семантическая близость, аналогия и синонимия. По результатам исследования видно, что синонимия уменьшается с уменьшением коэффициента аналогии, как показано на примерах выше. Первые 3-5 слов являются синонимами, но далее сходство с выбранным словом уже не настолько явное, а после 10-12-го слова встречаются только далекие аналоги по угадываемым предикатам (что общего можно сделать с исходным и данным словом). Мы делаем выводы о том, что аналог – это интуитивно ощущаемая нами близость слов по общим синтаксическим предикатам, а синоним – это сильный аналог по большему числу предикатов. И это определение, которое мы можем дать семантической аналогии на основании нашего исследования.

Проведенное исследование можно интерпретировать как экспериментальное доказательство того, что человек на интуитивном уровне «чувства аналога» ощущает сходство по множеству предикатов, но, так как их много, человек не может их назвать (все

сразу), чаще всего вместо этого называя самый распространенный из них. Например, для различных растений как аналогов распространенными предикатами будут «растут», «цветут», но это далеко не все общие предикаты разных растений. Это приоткрывает тайну нашей интуиции: мы не можем выразить одним словом множественные факторы семантической близости, а единичные не всегда обладают необходимой общностью для выборки тех предикатов, что являются общими для аналогов. Более того, данный метод применим с тематическим ограничением семантической близости, как показано выше, что устраняет омонимию и может применяться как способ получения аналогов в конкретном контексте. Этот фактор «интуиции сходства» отчасти объясняет, почему все подходы к семантическому описанию мира с помощью установки однозначных и единичных связей, как онтологии и семантические графы, не могут быть верными, что является важным следствием изученного метода поиска сходства по аналогии.

Одно из развитий метода – это разбивка слов на корни и аффиксы, чтобы создать большее число векторов предикатов, как между частями слов (приставка, корень, суффикс, окончание), так и между словами как членами предложения, как в предложенном методе (предикаты существительное-глагол, прилагательное-существительное и т.п.). Их сочетание создает еще большее пространство для возможных пересечений семантических свойств. Например, любое прибавление окончания -ик к слову, как известно, сразу создает новую семантику слова как уменьшительного. И тут, по нашему мнению, действует ровно такой же принцип сходства с такими же словами на -ик, как в предложенном методе, которые все обозначают маленькие предметы (не само окончание -ик создает уменьшение, а сходные слова на -ик) и по этому признаку являются аналогами. Этот принцип действует с ограничениями, прибавка -ик к чему-то априори большому не воспринимается как правильное, например, в слове гор-ик (исключение – материк). Но в целом этот известный эффект схож по принципу действия с исследованным нами методом аналогии.

Значение исследования для искусственного интеллекта

Исследованный метод открывает дискуссию о том, что результатов, которые были достижимы только нейросетевыми

подходами, можно добиться методами дистрибутивной семантики. И не только в аналогах – мы продолжаем работы по генерации текста на основе такого подхода. Это может открыть новый подход в методах искусственного интеллекта (ИИ), в частности поможет создать методы с обучением ИИ во время использования, так как PredicatIDF позволяет в интерактивном режиме добавлять новые предикаты в данные, на которых строятся аналогии. Появится возможность сразу создавать новые аналогии с выбранными предикатами. Вышеизложенное пока недостижимо в нейросетевом подходе, где обучение и использование разделены. Но не только в этом преимущество предлагаемого метода для использования в ИИ.

Мелани Митчер утверждает, что аналогия является ключевой способностью интеллекта [2, с. 145], поэтому данное исследование нам видится важным. Охарактеризованный метод определения аналогов может быть использован как альтернативный способ создания систем искусственного интеллекта, с обобщением по любому числу предикатов из контекста. LLM – это большой набор языковых шаблонов с небольшой аналогией по ним. Представляемый метод позволяет делать более глубокое обобщение шаблонов по аналогии. Практически любой факт, выраженный в виде клаузы, может стать шаблоном через аналогию его актантов. А так как и для аналогии по предикатам требуется совсем немного данных, то при создании систем искусственного интеллекта на основе предложенного метода потребуется на порядки меньшее число данных в отличие от нейросетевых систем. При этом аналогия может быть контекстно-зависимой и интерпретируемой – метод позволяет быстро определять, по каким предикатам аналогичны выбранные слова, что невозможно сказать по существующим способам аналогии, таким как word2vec и LLM.

Мы предлагаем называть такие факты с сильным обобщением по аналогии когнитивными схемами. Когнитивные схемы прочно заняли свое место в психологии интеллекта, но мало что известно о том, как они создаются и функционируют. Известно только, что схема – это некая общность разных случаев, аналогичных друг другу. Как, например, «все животные размножаются». Когнитивная схема формируется как устойчивая структура по мере пополнения аналогами первичного факта как примера, служащего ее прототипом. И те примеры, которые приобретают больше

аналогов с большей частотой встречаемости, становятся привычными шаблонами мышления за счет частотности их применения. И в них формируются прототипичные примеры как наилучшие выражения когнитивных схем. Это примеры с наиболее частотными предикатами, если для них нет обобщающих слов, что можно исследовать! В примере с «все животные размножаются» прототипом будут зайцы или свиньи как домашние животные, размножение которых происходило в деревне, где живут люди. Выделение когнитивных схем открывает перспективы развития методов комбинирования схем по общему контексту, например метафоры. Но данное утверждение требует отдельного исследования.

Заключение

Исследованный метод открывает интересные возможности и создает альтернативное направление развития искусственного интеллекта. Основные преимущества данного подхода заключаются в следующем:

1. Отсутствие необходимости обучения на огромном количестве данных.
2. Обучение во время использования, когда новые предикаты и схемы клауз сразу входят в базу знаний системы.
3. Интерпретируемость и управляемость таких систем с точки зрения используемых схем и аналогий при выводе следующей клаузы.

К трудностям относится то, что подобные системы работают только на уровне предикативной структуры, которую надо переводить в разговорную с учетом текущего диалога или запроса, что возможно сделать с помощью языковых моделей на основе трансформера, обученного только на уровне речевых шаблонов, без необходимости знания, для чего достаточно LLM с 1B параметром.

Выражаем надежду, что наша статья подтолкнет к исследованиям данного метода исследователей и совместно удастся развить его до новых основ ИИ.

Литература

1. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. <https://arxiv.org/abs/1301.3781>.
2. Митчелл М. Идиот или гений? Как работает и на что способен искусственный интеллект. М.: Издательство Corpus, 2022. С. 120-145.

- (Mitchell M. Idiot or genius? How does artificial intelligence work and what is it capable of. Publishing: Corpus House, 2022. P. 120-145).
3. Anna Rogers, Aleksandr Drozd, Bofang Li. The (too Many) Problems of Analogical Reasoning with Word Vectors. Available. January 2017. <https://www.researchgate.net/publication/318741605>.
 4. Aleksandr Drozd, Anna Rogers, Satoshi Matsuoka. Word Embeddings, Analogies, and Machine Learning. December 2016. <https://www.researchgate.net/publication/311843169>.
 5. Carl Allen, Timothy Hospedales. Analogies Explained: Towards Understanding Word Embeddings. <https://arxiv.org/abs/1901.09813>.
 6. Falcon Z. Dai, Word2vec Conjecture and A Limitative Result. <https://arxiv.org/abs/2010.12719>.
 7. Michael SC Thomas and Denis Mareschal. 1997. Connectionism and psychological notions of similarity. In The Proceedings of the 19th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum, Stanford, USA. P. 757-762. <http://eprints.bbk.ac.uk/4611>.
 8. Louis Fournier, Ewan Dunbar, Paraphrases do not explain word analogies. <https://arxiv.org/abs/2102.11749>.
 9. Tomáš Musil. Semantic Holism and Word Representations in Artificial Neural Networks. <https://arxiv.org/abs/2003.05522>.
 10. Tal Linzen. Issues in evaluating semantic spaces using word analogies. <https://arxiv.org/abs/1606.07736>.
 11. Paul Bartha. 2016. Analogy and analogical reason-ing. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy, Metaphysics ResearchLab, Stanford University. Winter 2016 edition. <https://plato.stanford.edu/archives/win2016/entries/reasoning-analogy>.
 12. Katrin Erk. What do you know about an alligator when you know the company it keeps. *Semantics and Pragmatics* 9(17):1-63. 2016. <https://semprag.org/index.php/sp/article/view/sp.9.17>.