Научная статья
УДК 004.048
https://doi.org/10.24143/2072-9502-2022-2-41-51

Применение методов искусственного интеллекта для решения задач поиска семантических ассоциатов на примере топонима «Москва»

Андрей Викторович Боровский[1], Елена Евгеньевна Раковская[2] *

[1, 2]Байкальский государственный университет
*Иркутск, Россия, rakovskaya19@mail.ru **

Аннотация. Актуальные проблемы топонимики подразумевают исследование отдельных слов с целью восстановления утраченного понятийного значения географических названий, выяснения того, как в них отразились характерные особенности рельефа местности, род деятельности населяющих ее людей и т. п. Цель исследования – определение происхождения топонима «Москва» с применением методов искусственного интеллекта. Применяется эмбеддинговая модель GeoWAC fastText на основе корпуса русскоязычных текстов сервиса RusVectōrēs для вычисления семантического сходства между словами. Модель предполагает определение семантических ассоциатов топонимов на основе векторного представления слов в семантическом пространстве и нахождение лексических векторов, наиболее близко расположенных к вектору исходного слова. Для анализа топонима применяются методы семантических ассоциатов, кластерный анализ, комбинированный метод, базирующийся на методе трансформации слова с утерянным смыслом и анализе семантических ассоциатов для множества трансформантов слова. Метод формализован применением модели, определяющей сходство исследуемого слова и ассоциатов, на основе разных вариантов модели для одного или нескольких корпусов текстов. Слова-ассоциаты, полученные искусственным интеллектом, рассматриваются как семантический кластер, вычисленное косинусное сходство между векторами – как мера сходства элементов в кластере. Для выявления различных гипотез возникновения топонима «Москва» проведен кластерный анализ совокупности первых десяти векторных ассоциатов для всех трансформантов этого слова. В результате выявлены четыре гипотезы: «знаменитый человек», «огнестрельное оружие», «пчеловодство», «кровососущие насекомые». Вычислены вероятности появления указанных гипотез на основе исследования частотности слов в корпусе языка. Основной является гипотеза «знаменитый человек».

Ключевые слова: эмбеддинговая модель, русский язык, метод трансформации слов, семантические ассоциаты, топоним «Москва», кластерный анализ

Для цитирования: *Боровский А. В., Раковская Е. Е.* Применение методов искусственного интеллекта для решения задач поиска семантических ассоциатов на примере топонима «Москва» // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2022. № 2. С. 41–51. https://doi.org/10.24143/2072-9502-2022-2-41-51.

Original article

Applying artificial intelligence methods for solving problems of searching for semantic associates: case of toponym Moskva

Andrey V. Borovsky[1], Elena E. Rakovskaya[2] *

[1, 2]Baikal State University
*Irkutsk, Russia, rakovskaya19@mail.ru **

Abstract. Actual problems of toponymy imply the study of individual words in order to restore the conceptual meaning of geographical names lost, to find out how they reflected the characteristic features of the terrain, the type of activity of the people inhabiting it, etc. The purpose of the study is to determine the origin of the toponym Moskva by using artificial intelligence methods. The GeoWAC fastText embedding model based on the corpus of Russian-language texts of the RusVecteres service is used to calculate semantic similarity between words. The model assumes defining the semantic associates of toponyms by using the vector representation of words in the semantic space and finding the lexical vectors most closely located to the vector of the original word. To analyze a toponym there is ap-

plied a methods of semantic associates, a cluster analysis, a combined method based on the method of transformation of a word with a lost meaning and the analysis of semantic associates for a set of word transformants. The method is formalized by using a model that determines the similarity of the studied word and associates based on different versions of the model for one or more text corpora. The associated words obtained by the artificial intelligence are considered as a semantic cluster, and the calculated cosine similarity between vectors is considered as a measure of the similarity of elements in the cluster. To identify various hypotheses of the origin of the toponym Moskva there has been carried out a cluster analysis of the totality of the first ten vector associates for all transformants of this word. As a result, four hypotheses were advanced: "a famous man", "firearms", "beekeeping", "blood-sucking insects". The probabilities of the occurrence of these hypotheses are based on the study of the frequency of words in the corpus of the language. The main hypothesis is a "famous person".

Keywords: embedding model, Russian language, method of word transformation, semantic associates, toponym Moskva, cluster analysis

For citation: Borovsky A. V., Rakovskaya E. E. Applying artificial intelligence methods for solving problems of searching for semantic associates: case of toponym Moskva. Vestnik of Astrakhan State Technical University. Series: Management, Computer Science and Informatics. 2022;2:41-51. (InRuss.) https://doi.org/10.24143/2072-9502-2022-2-41-51.

Введение

Развитие технологий искусственного интеллекта (ИИ) в сочетании с постоянно растущим объемом информации, увеличением производительности вычислительной техники, доступностью инструментария обработки данных привели к новым возможностям обработки естественного языка при проведении исследовательских работ. Для обработки лингвистической информации в последние годы успешно применяются эмбеддинговые модели [1, 2], полученные на основе корпусов текстовых данных. Эмбеддинги слов – семантически значимые векторные представления слов, в которых учитываются контексты употреблений слов в предложениях. Контекстно-зависимые модели могут быть полезны для изучения количественных характеристик естественного языка, для вычисления семантического сходства между двумя словами и для поиска слов, похожих на данное слово.

Применение эмбеддинговых моделей для определения семантических ассоциатов и смысловых значений слов, обозначающих названия географических объектов, является принципиально новым подходом для изучения происхождения топонимов [3].

Номинация географических объектов складывалась на протяжении длительного времени, поэтому в русском языке есть топонимы, появившиеся в разные исторические эпохи и связанные с разными сферами человеческой деятельности. Название каждого географического объекта отражает культурно-историческую информацию о народе, жившем на определенной территории, его верованиях, хозяйственной деятельности, этнических контактах, конкретных исторических событиях. Топонимы служат ценным материалом для исследования истории языка, поскольку доносят до нас слова, которые отсутствуют в каждодневной речи (утратили свое исконное значение) и существуют только в виде названий географических объектов. Изучение топонимики территорий как части историко-культурного наследия является важной и актуальной задачей.

Мотивы номинации топонимов очень разнообразны. Они определяются по естественно-географическим или топографическим условиям объекта, по связям с человеком и его деятельностью. Топонимы могут отображать мировоззрение народов, населяющих территорию, или относиться к языческим традициям. Иногда названия географических объектов имеют образный иносказательный смысл, который выражает эпическое, поэтическое творчество этноса, его характер и самобытность.

Естественно-географические условия объекта номинации определяются в виде характерных признаков объекта, особенностей природного ландшафта, непосредственных географических терминов (река, гора), видовых названий растений и животных (флора и фауна). Распространено отражение в названии исторически сложившегося рода деятельности местного населения. Часто связи географического объекта номинации с человеком отображаются в виде имен основателей поселений и городов или людей, оказавших большое влияние на духовную жизнь этноса (цари, вожди, духовные лидеры). Существуют топонимы, имеющие иноязычное, часто племенное происхождение. При этом с течением времени названия географических объектов могут видоизменяться, приобретая новые грамматические и фонетические признаки. Разобраться в многообразии происхождения топонимов бывает крайне сложно – первоначальный смысл названия географического объекта может быть утерян, исторические письменные источники о происхождении топонимов и мотивах номинации отсутствуют.

Статья посвящена исследованию на основе эмбеддинговых моделей одного важного для жителя Российской Федерации топонима – «Москва».

Постановка задачи и методы исследования

Целью исследования является определение происхождения топонима «Москва» на основе нахождения ИИ семантических ассоциатов к возможным трансформантам топонима. Для достижения поставленной цели решались следующие задачи:

– нахождение возможных трансформантов слова «Москва»;

– установление всех ассоциатов для каждого трансформанта с использованием программ ИИ на корпусе русского языка;

– семантическая интерпретация ассоциатов;

– формирование гипотез происхождения топонима «Москва» на основе кластерного анализа ассоциатов;

– расчет вероятности различных гипотез на основе частотности ассоциатов в кластерах.

Инструментальная часть исследования. В данной работе используется эмбеддинговая модель fastText [4] для русского языка, полученная на основе корпуса GeoWAC русскоязычных текстов (ресурс CommonCrawl), сбалансированных авторами разработки по географии Российской Федерации [5].

Параметры модели geowac lemmas none fast text skipgram 300 5 2020: корпус русского языка GeoWAC, размер корпуса – 2,1 млрд слов, объем словаря – 154 923 слов, средний частотный порог повторяемости слов не менее 150, алгоритм fast Text Skipgram (3–5-граммы), размерность вектора, содержащего ассоциаты, – 300, размер окна – 5 (количество слов в исходном термине), дата создания математического инструментария – октябрь 2020 г.

Тестирование метода проведено в исследовании русскоязычных топонимов Иркутской области с заранее известным смыслом [3]. Следует отметить, что ИИ для некоторых топонимов нашел совершенно неожиданные ассоциаты, которые заранее не были очевидны.

Сравнительно-исторический метод. Традиционно для определения происхождения слов, в том числе топонимов, применяется сравнительно-исторический метод. Этот метод предполагает рассмотрение лексического материала в развитии, с учетом неразрывной связи истории языка с историей его носителей. Метод базируется на изучении исторически сложившихся грамматических и фонетических закономерностей языка, диалектологии [6].

Метод семантических ассоциатов. Мотивы номинации топонимов, в том числе их явный или скрытый смысл, можно выявить при рассмотрении множества слов-ассоциатов, найденных с применением контекстно-зависимых моделей. Множество слов, которое охватывает определенную семантическую область и имеет структурированные отношения между элементами множества, определяют как семантическое поле [7]. Для определения происхождения названий географических объектов семантическое поле, состоящее из слов-ассоциатов, строится путем вычисления косинусного сходства между вектором топонима и векторами наиболее близко расположенных слов.

Теоретически векторные семантические представления слов и множества семантических ассоциатов дают возможность изучить слова по отношению к своим контекстам как в настоящее время, так и в разные исторические периоды. Можно использовать модели вложения слов для конкретных целей определения изменений значений слов или изучения дискурсивных пространств [8–12].

Семантические поля, характеризующие смысл и происхождение топонимов с помощью контекстно-зависимых моделей, рассматриваются:

1. С определением косинусного сходства между векторами слова в разных семантических пространствах, т. е. векторами, полученными на основе отличающихся корпусов текстов, $d^G\left(w_i^1, w_i^2\right) = \text{cossim}\left(w_i^1, w_i^2\right)$, где d^G – глобальная мера сходства; w_i^1 и w_i^2 – векторы слова i для семантического пространства 1 и 2; $\text{cossim}\left(w_i^1, w_i^2\right)$ – косинусное сходство векторов w_i^1 и w_i^2.

2. На основании списка слов-ассоциатов с вычислением косинусного сходства векторов главного слова и слов-ассоциатов, $s^1(i,j) = \text{cossim}\left(w_i^1, w_j^1\right)$, где $s^1(i,j)$ – набор слов-ассоциатов с вычисленным косинусным сходством для пространства 1; w_i^1 – вектор главного слова; w_j^1 – вектор слова-ассоциата.

3. Посредством сравнения списка слов-ассоциатов и косинусного сходства векторов главного слова и слов-ассоциатов, полученных в разных семантических пространствах (с применением разных моделей и отличающихся корпусов текстов), $d^L\left(w_i^1, w_i^2\right) = \text{cossim}\left(s_{i,j}^1, s_{i,j}^2\right)$, где $d^L\left(w_i^1, w_i^2\right)$ – мера локального сходства (изменения локальной окрестности слова i в пространстве 1 и 2); $s^1(i,j) = \text{cossim}\left(w_i^1, w_j^1\right)$ и $s^2(i,j) = \text{cossim}\left(w_i^2, w_j^2\right)$ – наборы слов-ассоциатов с вычисленным косинусным сходством для пространства 1 и 2.

Кластерный анализ. Для интеллектуального анализа лингвистической информации применяются методы кластеризации в семантических пространствах, отображающих смысловые характеристики слов естественного языка.

Семантические пространства строятся с использованием эмбеддинговых моделей. Слово представляется вектором в многомерном пространстве, который характеризуется направлением и длиной. Длина вектора есть частота употребления слова в рассматриваемом корпусе языка. Близкие друг к другу слова образуют лучевой кластер. Аналогом в физике является лучевая трубка. В исторической перспективе частотность некоторых слов в корпусе языка уменьшается, а некоторых увеличивается. Это означает, что часть слов забывается, а другая часть входит в употребление. Язык изменяется во времени.

Преимущества семантических пространств, генерируемых алгоритмами встраивания слов в эти

пространства, заключаются в том, что они обучены на больших массивах текстовой информации и могут отражать контекст использования, смысл слов, разнообразие и динамику естественных языков лучше, чем словари и лексические базы данных.

Кластеризация в семантических пространствах позволяет группировать слова, сходные по смыслу, в однородные группы (кластеры) для выявления структурных и семантических закономерностей в лингвистических данных. Рассматриваются связи между словами, имеющими похожие значения, прямые и переносные значения слов, семантические отношения с точки зрения диахронических изменений, стилистическая дифференциация слов. Формально слова-ассоциаты, полученные из распределенного векторного представления слов, можно рассматривать как семантический кластер, и вычисленное косинусное сходство между векторами – как меру сходства элементов в кластере.

При анализе кластеров ассоциатов можно идентифицировать оттенки значения и смысл слов, в том числе неявный. Полученные смысловые значения могут указывать на версии происхождения слова (топонима). Для изучения происхождения географических названий используются возможности кластерного анализа, связанные с выявлением информации о словах, записанных с грамматическими ошибками, о редких, устаревших словах. Полезная информация о семантических и логических связях топонимов с точки зрения их этимологии может быть получена при определении семантики топонимов по аналогии, при сравнении смыслов слов-ассоциатов кластера, а также при проведении математических вычислений с векторами слов.

Комбинированный метод. В статье применяется комбинированный метод, базирующийся на методе трансформации слов с потерянным смыслом [13] и анализе семантических ассоциатов для совокупности трансформантов слова, что дает возможность выявлять новые закономерности в трактовке смыслов топонимов. Примеры трансформации слов: отбрасывание окончания слова; замена глухих и звонких согласных (в-б-п, г-к-х, с-з-ж, с-ш, ч-щ); изменение гласных в корне слова (о-а, о-у, а-я). Целесообразность применения комбинированного метода определяется историческими факторами.

Естественный язык является отражением социокультурных, исторических, этнических отношений в обществе. Языковая ситуация в Древней Руси характеризуется следующим образом. С одной стороны, существует церковнославянский язык, так называемый «книжный» язык, или язык культуры, язык сакральный, на котором записаны все библейские и канонические тексты, и отдельно от этого языка существует древнерусский язык – язык повседневного общения. Со временем происходят процессы, которые можно охарактеризовать как взаимное влияние языков. Церковнославянский язык ассимилируется в русский национальный язык в литературных произведениях Ломоносова, Радищева, Державина и других литераторов XVIII в., который в дальнейшем преобразуется Пушкиным и Лермонтовым в русский литературный язык в XIX в. Русское влияние на церковнославянский язык проявляется в том, что некоторые языковые признаки усваивались церковнославянским языком в русской редакции [14].

Таким образом, имеются исторические документы, которые записаны на церковнославянском языке, или на языке элиты, и документы, которые записаны на языке разговорного общения, например путевые заметки казаков, участвовавших в военных походах. В этих документах орфографические нормы имеют неустойчивый характер или вообще отсутствуют. Вариативность орфографических норм дает возможность разного написания одного и того же слова, что определяется орфографическими и фонетическими традициями первоисточника текста (говором), применением скорописи для экономии писчего материала, наличием в тексте словоупотреблений живой речи [14].

Нахождение трансформантов топонима «Москва»

Мощным инструментом исследования терминов с потерянным смыслом является метод трансформации слов, описанный в разделе «Комбинированный метод». Мы впервые применим этот метод к исследованию ИИ важнейшего для русского человека топонима «Москва». Понятийный смысл топонима «Москва» в настоящее время утерян. Поставим задачу восстановить этот смысл в рамках корпуса русского языка GeoWAC с использованием программы ИИ fastText.

В истории России было принято давать названия новым городам, в том числе новым столицам, в честь знаменитых правителей страны, государственных деятелей и духовных лидеров: Петербург, Петроград, Ленинград, Екатеринбург, Екатеринослав, Екатеринодар, Днепропетровск, Сергиев-Посад, Пушкино, Лермонтов и т. д. Топоним «Москва», вероятно, не является исключением. С большой долей вероятности столица ранней России названа в честь великого полководца и церковного реформатора XV в., имевшего видоизмененные прозвища Мешех или Мосох, на что указывают трансформанты топонима «Москва».

Недавно была высказана версия арабского происхождения топонима «Москва», который на арабском языке означает просто «столица, столичный город». Даже если это так, то правитель, дававший название новому городу-крепости, использовал игру слов на арабском и старорусском языках.

Рассмотрим следующие виды трансформации слов: отбрасывание изменчивого окончания; замена глухих и звонких согласных (в-б-п, г-к-х, с-з-ж, с-ш, ч-щ...), изменение гласных в корне слова (о-а, о-у, а-я...). Для изучения топонима «Москва» этого будет достаточно.

Слово «Москва» содержит старое окончание «-ва». Сегодня его считают сочетанием суффикса «-в» и окончания «-а». В качестве примеров можно привести следующие старые русские слова (трава, ботва, тыква, крапива, плотва, бритва, канава и т. д.). Возможные виды трансформации топонима «Москва» представлены в табл. 1.

Таблица 1

Table 1

Варианты трансформации топонима «Москва» (моск – мск)

Уровень трансформации	Трансформанты топонима «Москва»			
1	мсх	мшк	мзк, мсг	мскн
	Мосох Моисей Христос	мошка мушка мишка	–	Моисей Князь
2	мшх	мшкт	мзг	мсхн
	Мешех	мушкет Мишка Т	мозг	Моисей Хан

Нахождение ассоциатов для трансформантов топонима «Москва» и их семантический анализ

При помощи модели GeoWAC fast Text получены семантические ассоциаты в том виде, в котором они приведены в табл. 2, цифры после ассоциатов представляют собой косинусы углов между многомерными векторами – исходным и ассоциатом – в используемой математической модели русского языка.

Таблица 2

Table2

Первые 10 семантических ассоциатов трансформантов топонима «Москва», полученных с применением модели GeoWAC fastText

The first 10 semantic associates of transformants of the toponym Moskva obtained by using the GeoWAC fastText model

Согласные буквы	Трансформанты топонима «Москва»	Семантические ассоциаты
Мск (0)	Москва	Санкт-Петербург 0,82; Петербург 0,80; Казань 0,79; Калининград 0,76; Тверь 0,74; Екатеринбург 0,74; Питер 0,74; москво 0,74; Краснодар 0,73; санкт 0,73
Мск (0)	моск	Москва 0,61; москво 0,57; Петербург 0,54; московский 0,53; московия 0,52; санкт 0,50; Питер 0,50; Санкт-Петербург 0,49; Саратов 0,47; Тверь 0,47
мшк (1)	мошк	мошка 0,70; мошкара 0,68; комар 0,59; насекомое 0,55; мошковский 0,52; таракан 0,52; мураво 0,51; слепень 0,51; кровососущий 0,50; москит 0,49
мшк (1)	мушк	мушка 0,58; ружье 0,42; гладкоствольный 0,41; охотничий 0,39; мушкетер 0,38; рогатка 0,38; карабин 0,38; арбалет 0,37; шпага 0,37; ружейный 0,37
мзг (2)	мозг	мозга 0,80; мозговой 0,75; мозги 0,72; мозод 0,72; нейрон 0,69; мозжечок 0,68; спинной 0,68; гипоталамус 0,67; надпочечник 0,63; спинномозговой 0,63
мшкт (2)	мушкет	сабля 0,65; револьвер 0,64; ружье 0,64; винтовка 0,64; арбалет 0,63; пушка 0,59; мушкетер 0,59; дробовик 0,58; гладкоствольный 0,58; шпага 0,58
мшкт (2)	Мишка Тверской	Тверской 0,60; вологодский 0,56; ярославский 0,54; пермский 0,53; мастерской 0,52; ташкентский 0,52; калужский 0,52; ивановский 0,51; воронежский 0,51; мурманский 0,51
мсх (1)	мосох	самосохранение 0,62; инстинкт 0,60; инстинктивный 0,56; посох 0,56; разум 0,54; инстинктивно 0,52; сатана 0,49; эмпатия 0,49; моисей 0,49; стадный 0,48
мсхн (2)	моисейхан	Моисей 0,65; царь 0,51; мухаммед 0,51; авраам 0,51; пророк 0,50; аббас 0,50; иудей 0,50; сулейман 0,50; хусейн 0,48; мухаммад 0,48
мсх (1)	Моисей- христос	Иисусхристос 0,77; господин иисусхристос 0,75; христос 0,73; моисей 0,72; иисус 0,69; христов 0,65; спаситель 0,65; иаков 0,63; иоанн креститель 0,63; господень 0,63
мшх (2)	мешех	Мешеть 0,58; хорошево 0,50; солотвино 0,49; каховка-нововоронцовка-новотроицкое-скадовск-щорупинск-чаплинка 0,48; рогачик-высокополье-геническ-голаяпристань-горностаевка-железный 0,47; виска-новгородка-новоархангельск-новомиргород-новоукраинка-ольшанка-онуфриевка-петрово-светловодск-ульяновка-устиновка 0,47; писаревка-глухов-конотоп-краснополье-кролевец-лебедин-липовый-долина-недригайлов-путивль-ромны-середина-буда-тростянец-шостка-ямполь 0,47; рог-кринички-магдалиновка-марганец-межевая-никополь-новомосковск-орджоникидзе-павлоград-першотравенск-петриковка-петропавловка-покровское-пятихатки-синельниково-соленое-софиевка-терновка-томаковка-царичанка-широкое-юрьевка 0,47; люботин 0,47; днестровский-беляевка-березовка-болград великий 0,47
мскн (1)	Моисей-князь	князь 0,73; моисей 0,73; авраам 0,65; иисус 0,63; пророк 0,63; иаков 0,63; христос 0,62; апостол 0,61; царь 0,61; иоанн креститель 0,61

Семантический анализ ассоциатов

Взятые для анализа варианты написания «Москва» и трансформант нулевого порядка (отсутствуют изменения в согласных) *моск*, с точки зрения изучения смысла топонима, интереса не представляют. Получены ассоциаты, определяющие современное значение слова – «крупный город и столица России».

При анализе трансформантов 1-го порядка (изменения в одной согласной букве (с-ш) *мошк* и *мушк* были получены семантические ряды (см. табл. 2), связанные с кровососущими насекомыми (*мошка*, *мошкара*, *комар*, *москит и др.*), добычей меда и изготовлением медовухи (*пчеловек*, *пчитывать*, *пчела*, *пчелиный*, *насекомое*, *муравo*, *пчеловод*, *улей*) и с огнестрельным оружием (*мушка*, *ружье*, *гладкоствольный*, *охотничий*, *мушкетер*, *карабин*, *ружейный*, *рогатка*). Слово «мушка» означает ружейный прицел. Слово «рогатка» возникло, поскольку мушкет при стрельбе ставили на рогатку.

Трансформант 2-го порядка *мшкт – мушкет*, в котором изменена одна буква *с-ш* и добавлена вторая *т*, имеет те же ассоциаты, что и трансформант *мушк*. Все они связаны с огнестрельным или холодным оружием.

Установить связь слова «мушкет» с именем князя *Мишка Тверской* (Михаил Тверской) не удалось. Программа на имя князя дала набор населенных пунктов, в каждом из которых, по-видимому, был известный человек с именем Мишка.

Ассоциаты для трансформанта 2-го порядка *мозг*, в котором изменены две согласные буквы *с-з* и *к-г*, связаны с устройством нервной системы человека. В XVI в., когда царь вместе с духовным владыкой давал название вновь отстроенной столице – Москва, само слово «мозг» было известно. Люди прекрасно знали, что значит *размозжить голову*, но устройство нервной системы человека вряд ли было известно. Большинство терминов, найденных ИИ, – из более позднего исторического периода.

Мосох – патриарх, описанный в Библии. Считается, что он привел свой народ на европейскую равнину, в междуречье Волги и Оки. Имя Мосох является трансформантом 1-го порядка от слова «Москва». Часть исследователей считает, что прозвище Мосох принадлежит 6-му (или 7-му) сыну одного из патриархов славян [15–23]. Авторы [24] считают Мосоха одним из сыновей царя-хана Татаро-Монголии начала XIV в. Возможно, что это прозвище сына Великого князя Бориса Александровича Тверского [13, с. 274–277]. В самом деле: Мосох = Мишех = Михаил. До сих пор существует славянское имя Миша, которое является сокращением от Михаила.

Мешех – другое имя Мосоха, является трансформантом 2-го порядка от слова «Москва». В ряду ассоциатов, найденных программой ИИ, первый ассоциат *мешеть*. Мы предлагаем продолжить цепочку трансформаций: *Мешех – мешеть – мечеть – мечтать – меч!*

Предположительно слово «мечеть» старорусского происхождения. На арабском языке мечеть называется совершенно по-другому. На латыни слово «мечеть» записывается как mosque, что практически совпадает со словом «Москва» [13]. Название «мечеть» отражает имя религиозного реформатора и полководца XV в. хана-первосвященника (халифа) Мешеха, он же Мосох, установившего правила поведения людей в районах, подверженных эпидемиям. Мечеть – это церковь, устроенная по правилам Мешеха.

В первой половине XV в. Татаро-Монголия ослабла, начали возвышаться Византия и Западная Европа. В южных областях разразились эпидемии чумы, холеры и оспы. В связи с этим царь-хан Татаро-Монголии развязал мировую войну, чтобы решить возникшие проблемы. Руководить войсками были назначены ханы Мосох и Иаков. Сведения об Иакове можно найти в [13]. Мосох был великим полководцем первой половины XV в. Он вел за собой на Запад южную армию Востока, поэтому летописцы Востока и Запада противоположно описывали Мосоха. Запад: при встрече с воинами Мосоха нужно было обладать чувством *самосохранения* (см. ассоциаты табл. 2), действовать *инстинктивно*, чтобы спастись. Мосоха отождествляли с *Сатаной* и *Змеем*-Горынычем (так люди воспринимали только что появившееся огнестрельное оружие, которому они ничего не могли противопоставить). Восток: у Мосоха был волшебный посох (*мушкет*), Мосох обладал великим *разумом*, он обладал чувством *эмпатии*, т. е. понимал других людей, Мосоха отождествляли с *Моисеем*, поскольку он, как Моисей, вел стадо военное и людское. Мосоха считали *чародеем*, *эльфом*, *жрецом*, и в самом деле Мосох был священником, *пророком*, поскольку он установил жесткие правила ведения военных действий, а также поведения солдат и обычных людей в зоне эпидемий, чем спас множество жизней. Мосоха путали с другим полководцем средневековой войны – святым *Иаковом*, который вел войска через Польшу, немецкие земли, Францию и Испанию, в то время как Мосох воевал на Кавказе, в Причерноморье, на Балканах и в Малой Азии.

Далее в колонке ассоциатов ИИ перечисляет последовательность населенных пунктов. Возможно, это походы Мешеха-Мосоха по территории Южной России, Болгарии и южной Польши или другие походы, которые ИИ посчитал связанными с именем Мешеха-Мосоха

Интересными являются ассоциаты, к которым приводят трансформанты 1-го и 2-го порядков с именем Моисей-хан, князь и Христос (имеется в виду, что по значению для людей деятельность Моисея похожа на деятельность Христа). Все они являются трансформантами топонима «Москва»!

Имеются библейские ассоциаты: *царь*, *пророк*, *иудей*. Имеются библейские имена: *Моисей*, *Мухаммед*, *Авраам*, *Аббас*, *Сулейман*, *Хусейн*, *Мухаммад*.

Исключительно интересными являются второе и последнее имена. Как показано в [13], имя Мухаммет образуется в результате небольшой трансформации латинскими летописцами рукописного русского имени Михаил Тверской с одной буквой фамилии на конце. Имена Мухаммад и Мехмет стоят в этом же ряду. Получается, что программа ИИ и метод трансформации слов связывают имена Мосох – Мешех– Моисей – Мухаммет – Мухаммад – Мехмет – Михаил Т (Михаил Тверской). В реальной истории существовал, по-видимому, один церковный реформатор и полководец XV в., имя которого было искажено латинскими летописцами при переписывании и размножено по различным летописным и церковным источникам.

Выскажем предположение, откуда появилось имя *Моисей*. Возможно, князь или хан-полководец завершал какие-то документы надписью: «Мы сей князь» (или хан). Через 50–100 лет летописец прочитал надпись как *Моисей князь* (хан). Далее имя в скорописи сократили до согласных букв *мск*. Отсюда в результате мог произойти топоним «Москва».

Авторы статьи считают, что имя *Михаил* старорусского происхождения. В самом деле: Михаил = = Ми-ка-ил = Мы как эл = мы как бог, что означает «божественный ребенок». Аналогичная трактовка имени существует у еврейского народа [25]: Михаи́л (ивр. מִיכָאֵל, Михае́ль) – мужское личное имя еврейского происхождения. Происходит от ивр. מי כמו אלוהים («ми кмо элохим», сокращенно «ми-ка-э́ль») – буквально «Кто как Бог».

Слово *ал=эл* на старорусском языке означало «бог» [26]. Выскажем следующую гипотезу: Аллах = = Ал-лах = Ал-рэх = бог-царь. До 1380 г. параллельно существовали две ветви христианства: царское христианство, в котором император объявлялся богом на земле, и апостольское христианство, которое объявляло последним богом на земле Иисуса Христа. В результате серии религиозных войн 1376–1402 гг. победило апостольское христианство. В православии слово «Аллах» не применяется, однако в русском языке слово *ал-эл* закрепилось в некоторых топонимах: Урал = Ур – ал = Бог Ур, Арал = Ар – ал = земля Бога, Марий Эл = Мария бого (родица).

Кластерный анализ ассоциатов

Вектор, составленный из ассоциатов, представляет собой лучевой кластер, однако работу, проделанную ИИ, следует сопроводить дополнительным анализом этого кластера. Во-первых, следует исключить из рассмотрения некоторые ложные ассоциаты. Во-вторых, целесообразно выделить подкластеры, отличающиеся различным смысловым значением. В-третьих, в некоторые подкластеры необходимо добавить слова, ускользнувшие от внимания ИИ. При анализе топонима «Москва» возникают следующие подкластеры: 4 подкластера со значением «знаменитый человек» (см. табл. 2) и подкластеры «огнестрельное оружие», «пчеловодство» и «кровососущие насекомые».

В подкластер «Мешех» включены найденные нами дополнительные ассоциаты – *мечеть*, *мечтать*, *меч*, *пешеход*. Слово «пешеход» имеет в русском языке неизвестную этимологию. Наше объяснение следующее. Правители средневековой Татаро-Монголии после побед XV в. установили для жителей правило, по которому каждый житель был обязан хотя бы один раз пройти по путям войск империи, т. е. осуществить хадж. Этот термин допускает трансформацию *Мешех-хадж = мшхд = пшхд = = пешеход*: рукописная «м» легко переходит в рукописную «п». Отсюда и произошло слово «пешеход». Точно так же произошло слово *Мосох = посох*.

Введем понятие вероятности появления слова в большом кластере. Для этого разделим частоту появления слова в корпусе русского языка *n* на частоту появления любого слова из большого кластера в корпусе русского языка *N*. Указанные вероятности приведены в 3-м столбце табл. 3.

Таблица 3

Table3

Кластеры ассоциатов топонима «Москва»

Clusters of associates of the toponym Moskva

Слова-ассоциаты	**Частота появления слова в корпусе, *n***	**Вероятность появления слова, *pr*(*n*) = *n* / *N***
Мосох		
самосохранение	3 815	0,005025145
инстинкт	19 293	0,025412879
инстинктивный	1 934	0,002547479
посох	4 128	0,005437431
разум	48 855	0,064352158
инстинктивно	3 975	0,005235899
сатана	8 181	0,010776072
эмпатия	1 953	0,002572506
моисей	5 894	0,007763619
стадный	1 308	0,001722907
Итого	*99 336*	*0,1308*

Окончание табл. 3

Ending of table 3

Слова-ассоциаты	**Частота появления слова в корпусе, *n***	**Вероятность появления слова, *pr*(*n*) = *n* / *N***
Мешех		
мешеть	278	0,000366184
мечеть	36 456	0,048020106
меч	51 865	0,068316952
мечтать	140 802	0,185465409
пешеход	43 262	0,056985018
Итого	*272 663*	*0,3591*
Моисей Хан		
царь	59 888	0,078884905
пророк	28 115	0,037033281
иудей	5 747	0,00756999
мухаммед	6 584	0,008672492
мухаммад	5 352	0,007049693
Итого	*105 686*	*0,1392*
Моисей князь		
князь	57 931	0,076307131
Иаков	3 275	0,004313854
Иисус	23 080	0,030401142
Христос	39 838	0,052474901
апостол	20 641	0,027188474
Итого	*144 765*	*0,1906*
мошк		
мошка	3190	0,004201891
мошкара	682	0,000898335
комар	16657	0,02194072
мураво	3621	0,004769607
слепень	650	0,000856185
кровососущий	679	0,000894384
москит	934	0,001230272
Итого	*26 413*	*0,0347*
пчела		
медоносный	772	0,001016884
пчеловек	5 908	0,00778206
пчитывать	1 631	0,002148365
пчела	11 639	0,015330975
пчеловод	3 849	0,005069931
улей	6 216	0,00818776
пчелиный	7 655	0,010083221
Итого	*37 670*	*0,0496*
мушк		
мушка	5 372	0,007076037
мушкет	832	0,001095916
ружье	23 788	0,031333725
гладкоствольный	1 899	0,002501376
охотничий	24 902	0,032801094
мушкетер	2 790	0,003675008
рогатка	2 739	0,003607831
карабин	9 006	0,011862768
ружейный	1 321	0,001740031
Итого	*72 649*	*0,0957*

Сумма всех вероятностей равна 1. Далее свяжем вероятность семантической гипотезы с вероятностью появления любого слова, принадлежащего подкластеру в корпусе русского языка. Гипотезу «знаменитый человек» формируют 4 подкластера из табл. 3: *знаменитый человек = Мосох + Мешех + + Моисей Хан + Моисей князь*. Вероятность этой гипотезы составит:

P (знаменитый человек) = 0,1308 + 0,3591 + + 0,1392 + 0,1907 = 0,8198 ≈ 82 %.

Семантические основы появления гипотез «огнестрельное оружие», «пчеловодство», «кровососущие насекомые» обсуждались в работе [3]. Вычисления, выполненные в данной работе, приводят к следующим вероятностям появления указанных гипотез: «знаменитый человек» : «огнестрельное оружие» : «пчеловодство» : «кровососущие насекомые» = 82 : 9,6 : 5,0 : 3,4 %.

Указанные гипотезы лежат в основе происхождения топонима «Москва».

Заключение

Для определения происхождения топонимов с утраченным смыслом впервые использованы ме-

тоды на основе эмбеддингов слов – эмбеддинговые модели для вычисления семантических ассоциатов Word2vec с архитектурой CBOW и Skip-gram; модель fastText, основанная на построении семантических векторов *N*-грамм слов. Преимуществом модели fastText является возможность работать с редкими и устаревшими словами. Анализ топонима «Москва» и его трансформантов в данной работе проводился с применением модели GeoWAC fastText русскоязычного корпуса GeoWAC (2,1 млрд слов), сбалансированного по географии России авторами разработки.

Новые результаты получены при изучении топонима «Москва». Предложено анализировать старинные топонимы с забытым смыслом, используя метод трансформации слова. Применение этого метода в сочетании с программой ИИ к топониму «Москва» привели к возникновению гипотез о происхождении термина: от имени полководца XV в.; от наименования огнестрельного оружия (мушкет); от пчеловодства и добычи меда (пчела-муха?); от кровососущих насекомых (мошка, мушка, муха). Трансформированными именами «знаменитого человека» – полководца-священника, реформатора церкви – являются Мешех, Мосох, Мухаммет, Мехмет.

Таким образом, для определения происхождения топонима целесообразно применить метод трансформации слова в сочетании с математическим моделированием трансформантов на основе эмбеддинговых моделей русского языка. Для определения различных гипотез возникновения топонима «Москва» был проведен кластерный анализ совокупности первых десяти векторных ассоциатов, присущих данному топониму. В результате были выявлены 4 гипотезы: *«знаменитый человек»*, *«огнестрельное оружие»*, *«пчеловодство»*, *«кровососущие насекомые»*.

Выбранный подход позволил вычислить вероятности появления указанных гипотез на основе исследования частотности появления слов, составляющих кластеры, в корпусе русского языка. Эти вероятности соотносятся как 82 : 9,6 : 5,0 : 3,4 %. Основной считается гипотеза «*знаменитый человек*», что неудивительно, т. к. в России было принято называть города и столицы в честь знаменитых правителей и духовных лидеров.

Список источников

1. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // International Conference on Learning Representations. Scottsdale, 2013. URL: https://arxiv.org/abs/1301.3781 (дата обращения: 12.09.2021).

2. *Goldberg Y., Levy O.* Word2vec Explained: Deriving Mikolov et al.'s Negative-sampling Word-Embedding Method // ArXiv. 2014. URL: https://arxiv.org/abs/1402.3722 (дата обращения: 12.09.2021).

3. *Боровский А. В., Раковская Е. Е.* Исследование топонимов Иркутской области с применением методов искусственного интеллекта // Изв. Байкал. гос. ун-та. 2021. Т. 32. № 3. С. 382–390.

4. *Bojanowski P., Grave E., Joulin A., Mikolov T.* Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. 2017. V. 5. N. 1. P. 135–146.

5. *RusVectōrēs:* семантические модели для русского языка. URL: https://rusvectores.org/ (дата обращения 12.09.2021).

6. *Русский* язык. Энциклопедия. М.: Дрофа, 1997. 703 с.

7. *Jurafsky D., Martin J.* Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Pearson, 2009. 615 p.

8. *Azarbonyad H., Dehghani M., Beelen K., Arkut A., Marx M., Kamps J.* Words are malleable: Computing semantic shifts in political and media discourse // Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017. P. 1509–1518.

9. *Hamilton W. L., Leskovec J., Jurafsky D.* Cultural shift or linguistic drift? comparing two computational measures of semantic change // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. V. 2016. URL: https://nlp.stanford.edu/pubs/hamilton2016cultural.pdf (дата обращения: 12.09.2021).

10. *Kenter T., Wevers M., Huijnen P., de Rijke M.* Ad Hoc monitoring of vocabulary shifts over time // Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, New York. 2015. P. 1191–1200.

11. *Orlikowski M., Hartung M., Cimiano P.* Learning diachronic analogies to analyze concept change // Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. 2018. P. 1–11.

12. *Recchia G., Jones E., Nulty P., Regan J., de Bolla P.* Tracing shifting conceptual vocabularies through time // European knowledge acquisition workshop. 2016. P. 19–28.

13. *Носовский Г. В., Фоменко А. Т.* Библейская Русь: в 4 т. М.: Римис. Т. 1. 496 с.

14. *Успенский Б. А.* История русского литературного языка (XI–XVII вв.): учеб. пособие. М.: Аспект Пресс, 2002. 560 с.

15. *Татищев В. Н.* История российская. М.: Директ-Медиа, 2012. Ч. 1. 997 с.

16. *Ломоносов М. В.* Древняя российская история от начала российского народа до кончины Великого Князя Ярослава Первого или до 1054 года. М.: Рипол Классик, 2013. 152 с.

17. *Тредиаковский В. К.* Три разсуждения о трех главнейших древностях российских (1749 г.) / адаптир. перелож. В. Н. Васильева. М.: Белые альвы, 2013. 216 с.

18. *Азимов А.* В начале. М.: Изд-во полит. лит-ры, 1989. 374 с.

19. *Православная* энциклопедия. М.: Православ. энцикл., 2009. Т. XX: Зверин в честь Покрова Пресвятой Богородицы женский монастырь – Иверия. 751 с.

20. *Еврейская* энциклопедия: в 16 т. СПб.: Общество для научныхъ еврейских изданий и изд-ва Брокгауз–Ефрон, 1908–1913. Т. 10. 952 с.

21. *Никифор А.* Иллюстрированная полная популярная библейская энциклопедия. М.: Проспект, 2018. 1565 с.

22. *Паламарчук П. Г.* Москва или Третий Рим? Восемнадцать очерков о русской истории и словесности. М.: Современник, 1991. 363 с.

23. *Токарев С. А.* Мифы народов мира. Энциклопедия. Электронное издание. URL: https://archive.org/details/Myths_of_the_Peoples_of_the_World_Encyclopedia_Electronic_publication_Tokarev_and_others_2008/page/n416 (дата обращения: 12.09.2021).

24. *Носовский Г. В., Фоменко А. Т.* Царский Рим в междуречье Оки и Волги. М.: АСТ, 2006. 800 с.

25. *Суперанская А. В.* Современный словарь личных имен. Сравнение. Происхождение. Написание. М.: Айрис Пресс, 2005. 384 с.

References

1. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations.* Scottsdale, 2013. Available at: https://arxiv.org/abs/1301.3781 (accessed: 12.09.2021).

2. Goldberg Y., Levy O. *Word2vec Explained: Deriving Mikolov et al.'s Negative- sampling Word-Embedding Method*. ArXiv. 2014. Available at: https://arxiv.org/abs/1402.3722 (accessed: 12.09.2021).

3. Borovskii A. V., Rakovskaia E. E. Issledovanie toponimov Irkutskoi oblasti s primeneniem metodov iskusstvennogo intellekta [Studying toponyms of Irkutsk region by using artificial intelligence methods]. *Izvestiia Baikal'skogo gosudarstvennogo universiteta*, 2021, vol. 32, no. 3, pp. 382-390.

4. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017, vol. 5, no. 1, pp. 135-146.

5. *RusVectōrēs: semanticheskie modeli dlia russkogo iazyka* [RusVectōrēs: semantic models for Russian language]. Available at: https://rusvectores.org/ (accessed: 12.09.2021).

6. Russkiiiazyk. *Entsiklopediia* [Russian language. Encyclopedia]. Moscow, Drofa Publ., 1997. 703 p.

7. Jurafsky D., Martin J. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, Pearson, 2009. 615 p

8. Azarbonyad H., Dehghani M., Beelen K., Arkut A., Marx M., Kamps J. Words are malleable: Computing semantic shifts in political and media discourse. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1509-1518.

9. Hamilton W. L., Leskovec J., Jurafsky D. Cultural shift or linguistic drift? comparing two computational measures of semantic change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.Conference on Empirical Methods in Natural Language Processing.* Vol. 2016. Available at: https://nlp.stanford.edu/pubs/hamilton2016cultural.pdf (accessed: 12.09.2021).

10. Kenter T., Wevers M., Huijnen P., de Rijke M. Ad Hoc monitoring of vocabulary shifts over time. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, New York.* 2015. Pp. 1191-1200.

11. Orlikowski M., Hartung M., Cimiano P. Learning diachronic analogies to analyze concept change. *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.* 2018. Pp. 1-11.

12. Recchia G., Jones E., Nulty P., Regan J., de Bolla P. Tracing shifting conceptual vocabularies through time. *European knowledge acquisition workshop*. 2016. Pp. 19-28.

13. Nosovskii G. V., Fomenko A. T. *Bibleiskaia Rus': v 4 t* [Biblical Russia: in 4 volumes]. Moscow, Rimis Publ., 2004. T. 1. 496 p.

14. Uspenskii B. A. *Istoriia russkogo literaturnogo iazyka (XI–XVII vv.): uchebnoe posobie* [History of Russian literary language (XI-XVII centuries): textbook]. Moscow, Aspekt Press, 2002. 560 p.

15. Tatishchev V. N. *Istoriia rossiiskaia* [History of Russia]. Moscow, Directmedia Publ., 2012. Part 1. 997 p.

16. Lomonosov M. V. *Drevniaia rossiiskaia istoriia ot nachala rossiiskogo naroda do konchiny Velikogo Kniazia Iaroslava Pervogo ili do 1054 goda* [Ancient Russian history from beginning of the Russian people to death of Grand Prince Yaroslav I, or up to 1054]. Moscow, Ripol Klassik Publ., 2013. 152 p.

17. Trediakovskii V. K. *Tri razsuzhdeniia o trekh glavneishikh drevnostiakh rossiiskikh (1749 g.)* [Three conclusions on three most important Russian antiquities (1749)]. Adaptirovannoe perelozhenie V. N. Vasil'eva. Moscow, Belye al'vy Publ., 2013. 216 p.

18. Azimov A. *V nachale* [In the beginning]. Moscow, Izd-vo polit. lit-ry, 1989. 374 p.

19. *Pravoslavnaia entsiklopediia* [Orthodox Encyclopedia]. Moscow, Pravoslav. entsikl., 2009. Vol. XX: Zverin v chest' Pokrova Presviatoi Bogoroditsy zhenskii monastyr' – Iveriia. 751 p.

20. *Evreiskaia entsiklopediia v 16 tomah* [Jewish Encyclopedia of Brockhaus and Efron. 16 volumes]. Saint-Petersburg, Obshchestvo dlia nauchnykh" evreiskikh izdanii i izd-va Brokgauz–Efron, 1908-1913. Vol. 10. 952 p.

21. Nikifor A. *Illiustrirovannaia polnaia populiarnaia bibleiskaia entsiklopediia* [Illustrated complete popular biblical encyclopedia]. Moscow, Prospekt Publ., 2018. 1565 p.

22. Palamarchuk P. G. *Moskva ili Tretii Rim? Vosemnadtsat' ocherkov o russkoi istorii i slovesnosti* [Moscow or the Third Rome? Eighteen essays on Russian history and literature]. Moscow, Sovremennik Publ., 1991. 363 p.

23. Tokarev S. A. *Mify narodov mira. Entsiklopediia. Elektronnoe izdanie* [Myths of the peoples of the world. Encyclopedia. Electronic edition]. Available at: https://archive.org/details/Myths_of_the_Peoples_of_the_World_Encyclopedia_Electronic_publication_Tokarev_and_others_2008/page/n416 (accessed: 12.09.2021).

24. Nosovskii G. V., Fomenko A. T. *Tsarskii Rim v mezhdurech'e Oki i Volgi* [Tsar's Rome in interfluve of the Oka and Volga rivers]. Moscow, AST Publ., 2006. 800 p.

25. *Sravnenie. Proiskhozhdenie. Napisanie* [Modern dictionary of personal names. Comparison. Origin. Writing]. Moscow, AirisPress, 2005. 384 p.

Статья поступила в редакцию 15.12.2021; одобрена после рецензирования 24.01.2022; принята к публикации 14.04.2022
The article is submitted 15.12.2021; approved after reviewing 24.01.2022; accepted for publication 14.04.2022

Информация об авторах / Information about the authors

Андрей Викторович Боровский – доктор физико-математических наук; профессор кафедры математических методов и цифровых технологий; Байкальский государственный университет; andrei-borovskii@mail.ru

Елена Евгеньевна Раковская – аспирант кафедры математических методов и цифровых технологий; Байкальский государственный университет; rakovskaya19@mail.ru

Andrei V. Borovsky – Doctor of Physical and Mathematical Sciences; Professor of the Department of Mathematical Methods and Digital Technologies; Baikal State University; andrei-borovskii@mail.ru

Elena E. Rakovskaya – Postgraduate Student of the Department of Mathematical Methods and Digital Technologies; Baikal State University; rakovskaya19@mail.ru