УДК 004.85

МЕТОДЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ НАУЧНЫХ ИССЛЕДОВАНИЙ В ГЕОЛОГИИ

М. И. Патук[1] [0000-0003-3036-2275], В. В. Наумова[2] [0000-0002-3001-1638]

ФГБУН Государственный геологический музей им. В.И. Вернадского Российской академии наук, Москва, Россия

[1]patuk@mail.ru, [2]naumova_new@mail.ru

Аннотация

Приведен краткий обзор некоторых методов искусственного интеллекта в области наук о Земле. Отмечены перспективы применения указанных методов для получения новых знаний. Приведены результаты первых попыток авторов в применении методов обработки естественного языка для обработки научных статей по геологии. Обсуждены возможности развития работ в этом направлении.

***Ключевые слова**:* Искусственный интеллект, машинное обучение, обработка естественного языка, геология.

ВВЕДЕНИЕ

Известно много определений термина искусственный интеллект (ИИ). Например:

- ИИ — общий термин, описывающий системы, выполняющие когнитивные, познавательные функции, например, решение тематических проблем [1].
- ИИ – способность системы правильно интерпретировать внешние данные, извлекать уроки из таких данных и использовать полученные знания для достижения конкретных целей и задач посредством гибкой адаптации [2].

Уровень познания, необходимый для выполнения определенной задачи, определяется ее характером, поэтому рассматриваемый термин можно применять в отношении любого процесса поиска решения или интерпретации данных с использованием компьютера. Таким образом, понятие «искусственный интеллект» охватывает широкий спектр процессов, используется в контексте программного обеспечения и соответствующих услуг, в том числе связанных с машинным

обучением. В области ИИ существует много подходов и направлений. Мы остановимся на трех их них: экспертные системы, обработка изображений и обработка естественного языка (Рис. 1).

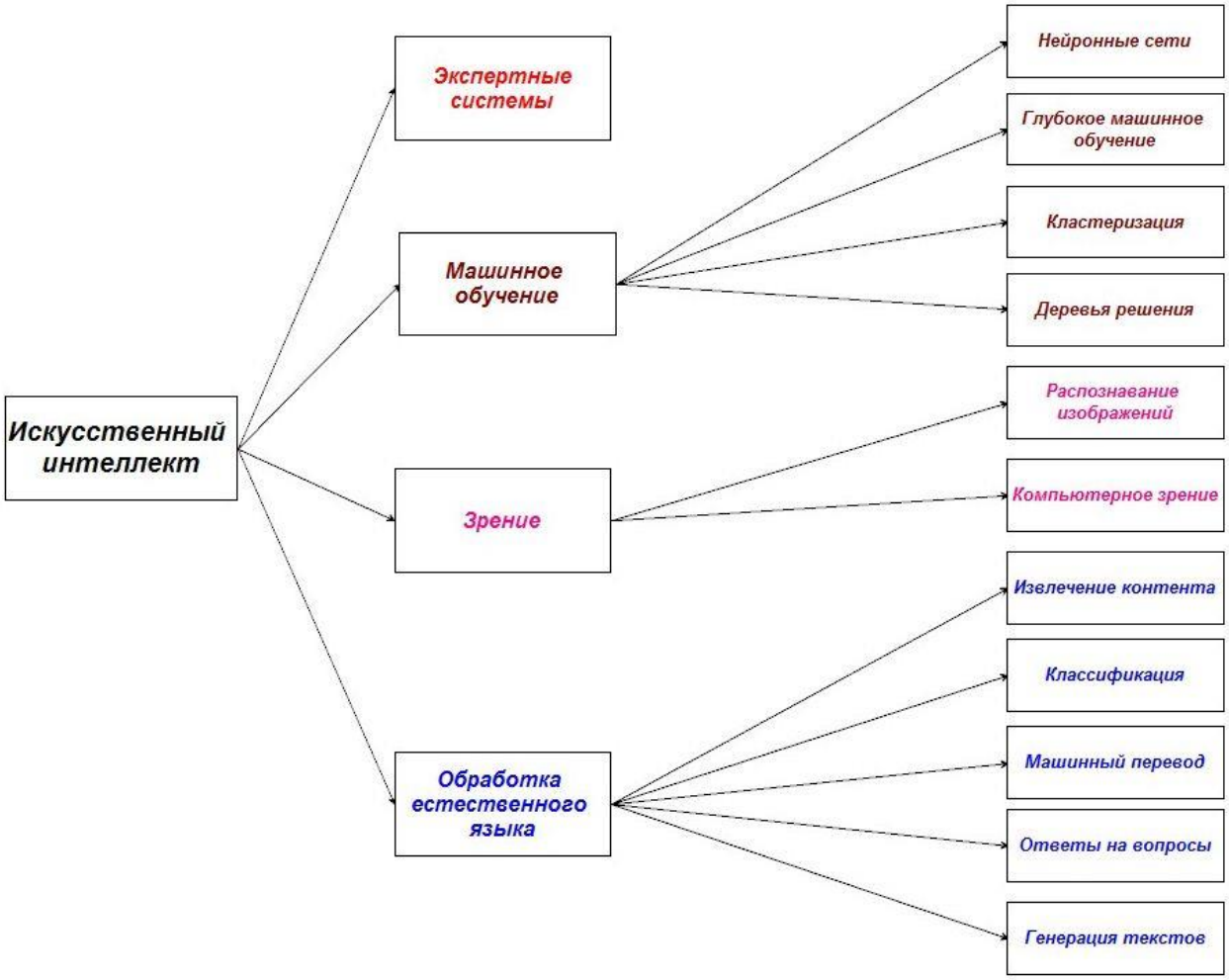

Рис. 1. Обозреваемые направления искусственного интеллекта

Экспертные системы в геологии

Экспертная система (ЭС) – это интеллектуальная программа, способная делать логические выводы на основании знаний в конкретной предметной области и обеспечивающая решение специфических задач. Для этого ее необходимо наделить функциями, позволяющими решать задачи, которые в отсутствие эксперта (специалиста в данной конкретной предметной области) невозможно решить правильно.

В области геологии одной из первых является экспертная система

PROSPECTOR [3], разработанная для оказания помощи геологам-поисковикам. Эта программа создана компаниями ESRI [4] (совместно с консультантами по геологии) и USGS [5]. Она способна давать три типа «советов»: оценку местности на предмет существования определенных залежей; оценку прогнозных ресурсов региона и выбор перспективных участков для бурения. Эксперт-геолог применяет очень ограниченные модели при выявлении территорий с возможными запасами, например, золота. В ЭС PROSPECTOR эти модели в закодированном виде находятся в памяти компьютера, они определенным образом интерпретируются при оценке того или иного участка. Одна из главных особенностей знаний экспертов-геологов заключается в том, что они неполны и неопределенны. В связи с этим используются специальные технические приемы, отличные от тех, что применяются в других (более определенных) экспертных системах.

Экспертная система «ОЛОВО» [6]

Для работы с ЭС «Олово» не требуется специальной подготовки картографического или иного материала. Единственное условие – квалифицированное владение пользователем совокупностью сведений, относящихся к комплексной многоуровневой характеристике объекта прогноза.

Экспертная система «Олово» воспринимает новые данные для решений одной или комплекса прогнозных задач по какому-либо конкретному объекту или участку территории в виде геологических знаний пользователя о данном объекте. Система непрерывно запрашивает пользователя, пока не будут заполнены все отсутствующие компоненты новой структуры (модели экзаменуемого объекта).

В процессе усвоения знаний об экзаменуемом объекте на каждом шаге развития диалога с пользователем происходят поэтапное формирование модели этого объекта, а также поэтапное последовательное сопоставление всех ее элементов с соответствующими элементами знаний, заложенных в базу знаний (БЗ) экспертных моделей.

Непосредственное решение прогнозных задач выполняется в системе с использованием нескольких, различных по своему смысловому содержанию методических приемов, каждый из которых обеспечен соответствующим математическим аппаратом. Один из таких приемов основан на методе аналогий. При его реализации за каждым шагом усвоения новых знаний об экзаменуемом объекте

следует определение вероятной схожести (в численном выражении) конструируемой модели объекта с той или иной экспертной моделью. В процессе диалога с пользователем по мере накопления новых знаний эти вероятности могут изменяться. Все текущие изменения учитываются при формировании прогнозного заключения, в котором указывается степень соответствия данного объекта определенной экспертной модели.

Другой методический прием содержит в своей основе метод распознавания образов. Эталонные образы в виде экспертных моделей заложены в БЗ. Поэтапный анализ положения конкретных признаков экзаменуемого объекта в многомерном пространстве экспертных признаков позволяет системе производить классификацию знаний, получаемых в процессе диалога с пользователем. Наличие в системе специфических решающих правил обеспечивает возможность использования в процессе формирования прогнозного заключения наиболее значимых признаков с оценкой степени их влияния на окончательные выводы и демонстрировать пользователю численное значение вероятностей распознавания образа. Таким путем достигается дифференциация окончательных результатов решения задачи по степени их надежности.

Таким образом, ЭС «Олово» позволяет решать задачи прогнозирования на стадии регионального изучения недр. Более локальные прогнозы, касающиеся количественной оценки ресурсов оловорудного узла, уровня эрозионного среза объекта и возможности обнаружения промышленных скоплений руд решается с меньшей долей вероятности. Это связано со спецификой созданной базы знаний, которая включает широкий спектр разнообразной геологической информации, но лишена знаний экономического характера. Тем не менее, по ряду косвенных характеристик и различных критериев (минералого-геохимических, геофизических, геолого-минералогических и др.) эта система дает количественные прогнозы и определяет вероятность их подтверждения.

Сейчас количество экспертных систем исчисляется тысячами и десятками тысяч. В развитых зарубежных странах сотни фирм занимаются их разработкой и внедрением.

SOLSA [7] – первая автоматизированная экспертная система для анализа

керна на месте. Благодаря доступу к данным в режиме онлайн ожидается значительная экономия на количестве буровых скважин, точности геомоделей и экономической оценке запасов руды. ЭС SOLSA идеально отвечает потребности в «Новых технологиях устойчивой разведки и геомоделей» SC5-11d-2015. Целью ее создания была «разработка новых или усовершенствованных высокоэффективных и рентабельных, устойчивых технологий разведки», включая:

- комплексное бурение, оптимизированное для работы в сложных латеритных условиях со сложной смесью твердых и мягких пород, распространяемое также на другие типы руд,
- полностью автоматизированный сканер и анализатор фазовой идентификации, а также программное обеспечение, которое можно использовать и в других отраслях.

ЭС SOLSA впервые объединила неразрушающие датчики рентгеновской флуоресценции, рентгеновской дифракции, колебательной спектроскопии, 3D- и гиперспектральной визуализации вдоль керна скважины. Для этой цели SOLSA разработала инновационное, удобное для пользователя и интеллектуальное программное обеспечение на уровнях TRL 4-6. Чтобы минимизировать риск и извлечь выгоду из новейших технологий, на рынке миниатюрных датчиков были выбраны подсистемы для аппаратного обеспечения.

ЭС SOLSA призвана произвести революцию в геологической разведке, сократив ее время на 50%, а время анализа – с 3–6 месяцев до реального времени и, таким образом, снизить воздействие на окружающую среду.

Машинное обучение в геологии

Машинное обучение — это класс количественных методов (под которыми зачастую понимают алгоритмы), предназначенных для ускорения процесса прогнозирования определенных показателей на основе некоторого прецедента [1]. В отличие от остальных направлений в ИИ, машинное обучение не требует ручного ввода в алгоритм правил принятия решений — они автоматически определяются системой по эмпирическим данным.

Существует широкий спектр алгоритмов машинного обучения, подходящих для выполнения специализированного геологического анализа. Исходный мате-

риал для их обучения обычно либо уже имеется, либо может быть получен самостоятельно. Таким образом, машинное обучение можно использовать с целью выявления геологоразведочных объектов в условиях избытка данных (например, решения Goldspot Discoveries [8], SRK Consulting [9]), автоматического выявления геологических зон залегания полезных ископаемых (Maptek [10]), оценки твердости руды на основе результатов анализа (неопубликованные работы), распознавания частиц золота по фотоснимкам пробы ледниковых отложений (IOS Services Geoscientifiques [11]).

В последние десятилетия наблюдается стремительный рост интереса к нейронным сетям, которые успешно применяются в различных областях — бизнесе, медицине, технике, геологии, физике. Нейронные сети используются всюду, где нужно решать задачи прогнозирования, классификации, нелинейной регрессии или управления. Такой впечатляющий успех нейронных сетей определяется богатыми возможностями и простотой в использовании. Особенность работы нейросетей состоит в том, что такая сеть обучается на исторических данных, находит специфические паттерны, указывающие на зависимости внутри данных, и на их основе строит свой прогноз.

Одной из разновидностей нейронных сетей, предназначенной для обработки изображений и других точечных форматов, являются сверточные нейронные сети. В геологоразведке они применяются для выявления объектов (например, решения Orefox [12]), обработки и интерпретации сейсмических данных (Geolearn [13]), определения минералов-индикаторов в пробах ледниковых отложений (IOS Services Geoscientifiques [11]), а также количественного и качественного описания буровых кернов по их фотоснимкам (Geolearn [13]) или гиперспектральным данным (Solve Geosolutions [14]).

Последовательность входных данных анализируется с помощью такой разновидности нейронных сетей, как рекуррентные нейронные сети. Они адаптированы для анализа временных наборов данных, таких как временные последовательности или текстовая информация. В геологоразведке рекуррентные нейронные сети используют для выявления перспективных участков на основе отчетов, находящихся в свободном доступе (например, решения Goldspot Discoveries [8]),

или для геологического документирования данных бурения на основании измерений физических свойств пород (CGG).

Методы машинного обучения все чаще используются в горнодобывающей промышленности. Они эффективны в решении повторяющихся задач или задач с большим количеством многомерных данных (качественных и правильно обработанных) [1].

Объективность, продуктивность и адаптивность алгоритмов машинного обучения делают их идеальным решением широкого спектра проблем различного масштаба. Однако подготовка и внедрение таких технологий в разведке и добыче требуют немалого опыта. Моделирование — это комплексная работа, которой сопутствуют характерные сложности, и качество входных данных — не самая последняя из них.

Анализ изображений в геологии

В геологии большую часть времени работы исследователей занимают визуальная диагностика и описание горных пород. Состав породы, ее структура и текстура должны быть определены и описаны. Это занимает много времени, и естественно возникает желание автоматизировать эту работу. В работе [15] предложен новый подход к автоматической классификации пород при описании кернов. Описаны доступные методы автоматической классификации пород, предложен новый подход применения сверточной нейронной сети. Поскольку для тренировки нейронной сети требуется большое количество данных, были предприняты специальные усилия для генерации дополнительных изображений. Приведено описание и проведено сравнение разных архитектур нейронной сети. Была достигнута точность 72%. Система автоматической классификации натренирована на 20 000 изображений образцов из 10 нефтяных и газовых полей различных геологических условий. Описаны ограничения применения полученной модели.

Получить достаточное количество изображений для анализа не всегда возможно. Поэтому в работе [16] авторы воспользовались специальными методами обучения с нулевым результатом (zero-shot learning) и обучение с малым числом примеров (few-shot learning). В результате стало возможным предложить открытый критерий для распознавания необработанных минералов, которые отсут-

ствуют в обучающей выборке. Также предоставлены дополнительные наборы образцов для сегментации, определения размеров образцов и классификации с малым числом примеров. Для всех описанных задач компьютерного зрения предоставлены базовые алгоритмы. В статье показана важность унифицированных данных для корректной работы распознавания минералов. Созданные коллекции изображений минералов выложены в открытый доступ для использования всеми желающими.

Обработка естественного языка в геологии

Обработка естественного языка — это общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза текстов на естественном языке [17].

Рассмотрим некоторые определения из данной области: языковой моделью называется набор свойств и методов по построению распределения вероятностей последовательности слов [18]. Языковая модель создает вероятности в процессе обучения на корпусе текстов. Корпусом текстов называется набор текстов, подобранный и обработанный по определенным правилам. В лингвистике, откуда пришел этот термин, под корпусом текстов понимается коллекция текстов, размеченная с помощью специальных тэгов. В области машинного обучения под корпусом текстов часто понимается просто их коллекция, без специальной разметки. Векторное представление слова (word embeddings) – назначение слову числового вектора на основе его анализа языковой моделью.

В области наук о Земле с помощью методов обработки естественного языка (NLP) решаются следующие задачи: выделение геологических и географических именованных сущностей (NER), извлечение пространственных и временных взаимосвязей, классификация, кластеризация, реферирование геологических отчетов и публикаций, ответы на вопросы.

Большие языковые модели, такие как BERT, ChatGPT и GPT-4, достигли большого успеха в применении к общеупотребительным языковым областям, таким как новостные, социокультурные, медийные. В области наук о Земле их применение не столь впечатляющее, в первую очередь, из-за отсутствия в их арсенале спе-

цифической геологической терминологии, поскольку она отсутствует в тех языковых корпусах текста, на которых обучались данные модели [19]. Авторами была создана большая языковая модель для наук о Земле – K2 [20] – на основе обучения GPT подобной языковой модели LLaMA. Для целей тренировки модели создан большой набор текстов, содержащий более 6 млн. записей, включающих статьи, метаданные статей и данные из Wikipedia. K2 – генеративная модель. Она может создавать тексты по теме наук о Земле и отвечать на соответствующие вопросы.

Предсказательные возможности методов обработки естественного языка основаны на статистическом языковом моделировании, имеющем в своей основе векторное представление слов (word embeddings). Векторное представление каждого слова создается на основе частот взаимного расположения других слов, находящихся вблизи выбранного слова в обучающем корпусе текстов. Таким образом создается контекстно-зависимое представление слов, что очень полезно для слов с множественными значениями (полисемия). Например, такие геологические термины, как щит, плита, кора, пояс, узел, чехол, трубка, осадки, мел (сокращенное от «меловой период»), мантия, свита, имеют значения, совсем отличные от общеупотребительных. Понять их геологический смысл можно только по контексту. Появление таких контекстно-зависимых моделей, как BERT, позволило значительно улучшить возможности этого класса моделей в обработке естественного языка.

Указанная контекстная зависимость современного поколения языковых моделей имеет свою обратную сторону. Чтобы получать адекватные результаты, необходимо тренировать такие модели на текстах, специфичных для каждой предметной области. В [21] дано сравнение результатов выполнения разными языковыми моделями тестов со специфической геологической терминологией. Модели, обученные на корпусах геологических текстов, превзошли модели, обученные на значительно больших корпусах общих текстов.

Указанная работа [21] делает упор на обучении двух моделей: GloVe – не контекстно зависимой и BERT – контекстно зависимой, на неразмеченных коллекциях текстов (геологические отчеты и научные публикации, доступные в свободном доступе). Созданы внутренние критерии оценки моделей (аналогии, образо-

вание кластеров, родство и ближайшее окружение) взамен внешних тестов. Показана возможность извлечения, с помощью указанных моделей, данных о геохимических и минеральных ассоциациях из необработанных текстовых данных геологической направленности.

В работе [22] выполнены извлечение ключевых слов и их визуализация (облако тэгов, индексы центральности) из геологических отчетов. Извлечение ключевых слов выполнялось на основе модифицированного алгоритма TF-IDF.

В работе [23] создана языковая модель для области наук о Земле (GeoVec), обученная на 280 000 научных статей из данной области. Для внутренней оценки модели были проведены тесты по созданию аналогий, на определение родственных терминов и деление терминов на категории. Было показано превосходство данной модели для области наук о Земле над стандартными моделями, которые обучались на общемедийных текстах.

Обработка естественного языка была использована для извлечения текстовых данных из описательной части геологических карт [24]. Выбирались данные описания пород, геологического возраста, литостратиграфические описания. Извлеченные данные преобразовывались в векторную форму, и с использованием статистических методов находились семантические связи между типами пород. Кроме того, с помощью тех же методов выполнялось предсказание территорий, перспективных на Zn-Pb оруденение.

В работе [25] обработка естественного языка была использована для классификации и 3-х мерного литологического картирования. Были использованы текстовые данные материалов бурения. Векторные представления слов были получены из заранее обученной модели GloVe [23]. Текстовые описания были размечены экспертами. Это позволило выполнить классификацию текстовых данных с помощью нейронной сети и посредством интерполяции создать приемлемые 3-х мерные литологические карты исследованного района Австралии.

Большой объем неструктурированной геологической информации содержится в геологических отчетах и научных статьях. Невозможно простым прочтением охватить эту лавину информации. Отсюда вытекает задача извлечения краткого резюме из имеющейся информации для быстрого первичного анализа имеющихся источников [26]. Геологические тексты обладают большой спецификой

из-за использования большого количества специфических терминов и стоящих за ними взаимосвязей и геологических концепций. Авторы предложили последовательный подход по извлечению таких терминов в виде геологических именованных сущностей на неразмеченных текстах геологических отчетов.

Задачи бинарной классификации

Из приведенного выше краткого обзора работ по применению методов обработки естественного языка в области наук о Земле видно, что для работы с этими методами необходимо иметь обученную языковую модель. И не просто обученную, а обученную на текстах из интересующей нас предметной области. Русскоязычные языковые модели существуют, но нам не удалось найти ни одной, обученной на текстах геологической направленности. Таким образом, для начала работы в этой области необходимо получить коллекцию геологических текстов.

Эта проблема оказалась решаемой, т. к. в Государственном геологическом музее им. В.И. Вернадского имеются два текстовых ресурса: архив научных публикаций с тематикой «Науки о Земле» (https://repository.geologyscience.ru/) [27] и wiki-Геология России (http://wiki.geologyscience.ru) [28]. Текстовые данные этих ресурсов находятся в SQL-базах, что позволяет достаточно легко их извлекать.

Первая задача, которая нам показалась интересной не столько с практической, но и с методологической точки зрения, – это бинарная классификация. Эта задача подробно и многократно описана в интернете. Мы выбрали следующее описание [29], как очень подробное и предоставляющее возможность скачать все обсуждаемые примеры с GitHub [30]. В упомянутой статье представлено 8 моделей для классификации. Последние две модели из списка мы не использовали, поскольку там используется заранее натренированная англоязычная модель. Естественно, мы не стали использовать предоставляемые данные, поскольку они не нашей тематики и к тому же англоязычные. Нами были выбраны 50 научных статей из архива публикаций (https://repository.geologyscience.ru/), касающихся месторождений золота и железа. PDF-файлы этих статей были конвертированы в текст с помощью пакета PDFReader [31]. Дополнительно тексты были очищены от всех некириллических символов и цифр с помощью регулярных выражений и удалены стоп-слова (слова типа предлогов, местоимений, которые не несут смысловой нагрузки).

Для дальнейшей обработки текст необходимо его токенизировать, т. е. разбить на отдельные слова и привести их в нормальную форму (единственное число, именительный падеж, мужской род). Для этого мы использовали две русскоязычные модели: 1-я модель – ru_core_news_lg [32]– автор Александр Кукушкин, 2-я модель – spacy-stanza [33], предыдущее название – StanfordNLP. Анализ текстов после токенизации показал, что модель spacy-stanza работает в нашем случае лучше, но с ошибками, коверкает некоторые слова, например, вместо «кристаллический» вставляет «каллический».

В соответствии с рекомендациями по подготовке исходных данных [34] все тексты после токенизации были вычитаны и отредактированы. Расчет по 6-ти указанным выше моделям оказался неудовлетворительным – низкая точность, высокие потери (Рис. 2).

```
Epoch 1/5

1/1 [==============================] - ETA: 0s - loss: 0.6754 - accuracy: 0.6800
1/1 [==============================] - 2s 2s/step - loss: 0.6754 - accuracy: 0.6800 - val_loss: 0.6609 - val_accuracy: 0.6667
Epoch 2/5

1/1 [==============================] - ETA: 0s - loss: 0.6322 - accuracy: 0.7600
1/1 [==============================] - 0s 172ms/step - loss: 0.6322 - accuracy: 0.7600 - val_loss: 0.6537 - val_accuracy: 0.6667
Epoch 3/5

1/1 [==============================] - ETA: 0s - loss: 0.5961 - accuracy: 1.0000
1/1 [==============================] - 0s 156ms/step - loss: 0.5961 - accuracy: 1.0000 - val_loss: 0.6470 - val_accuracy: 0.6667
Epoch 4/5

1/1 [==============================] - ETA: 0s - loss: 0.5652 - accuracy: 1.0000
1/1 [==============================] - 0s 156ms/step - loss: 0.5652 - accuracy: 1.0000 - val_loss: 0.6401 - val_accuracy: 0.6667
Epoch 5/5

1/1 [==============================] - ETA: 0s - loss: 0.5366 - accuracy: 1.0000
1/1 [==============================] - 0s 172ms/step - loss: 0.5366 - accuracy: 1.0000 - val_loss: 0.6330 - val_accuracy: 0.6667
```

Рис. 2. Результат обучения сверточной нейронной сети на малом числе примеров

После этого мы изменили свой подход к получению исходных данных: вместо полнотекстовых статей из архива публикаций (https://repository.geologyscience.ru) были выбраны абстракты статей, касающихся описания месторождений золота и железа. Всего было выбрано 1750 записей (1150 – про золото и 600 – про железо). Средняя длина строк текста – 120 слов. На этих данных мы получили более вдохновляющие результаты – точность достигла 92% (Рис. 3). Наилучшие результаты показала модель одномерной сверточной нейронной сети (1D Convolutional Neural Network). При этом мы не очищали эти тексты и не удаляли стоп-слова. Дополнительная чистка текстов и удаление стоп слов незначительно повысили точность (менее чем на 1%). Завершающий

этап – проверка полученной модели. На этом этапе на вход модели подается текст, который она раньше «не видела». Ее задача – отнести этот текст к одной из двух категорий, на которых она обучалась. Модель в основном успешно классифицировала тексты, как описывающие месторождения золота или железа. Этот результат с очевидностью показал, что при обучении языковых моделей количество (1750 против 50) имеет первостепенное значение.

```
28/50 [===============>..............] - ETA: 7s - loss: 0.0262 - accuracy: 1.0000
29/50 [================>.............] - ETA: 6s - loss: 0.0259 - accuracy: 1.0000
30/50 [================>.............] - ETA: 6s - loss: 0.0260 - accuracy: 1.0000
31/50 [=================>............] - ETA: 6s - loss: 0.0257 - accuracy: 1.0000
32/50 [=================>............] - ETA: 5s - loss: 0.0257 - accuracy: 1.0000
33/50 [=================>............] - ETA: 5s - loss: 0.0256 - accuracy: 1.0000
34/50 [==================>...........] - ETA: 5s - loss: 0.0256 - accuracy: 1.0000
35/50 [==================>...........] - ETA: 4s - loss: 0.0253 - accuracy: 1.0000
36/50 [===================>..........] - ETA: 4s - loss: 0.0253 - accuracy: 1.0000
37/50 [===================>..........] - ETA: 4s - loss: 0.0253 - accuracy: 1.0000
38/50 [===================>..........] - ETA: 3s - loss: 0.0253 - accuracy: 1.0000
39/50 [====================>.........] - ETA: 3s - loss: 0.0251 - accuracy: 1.0000
40/50 [====================>.........] - ETA: 3s - loss: 0.0250 - accuracy: 1.0000
41/50 [=====================>........] - ETA: 2s - loss: 0.0250 - accuracy: 1.0000
42/50 [=====================>........] - ETA: 2s - loss: 0.0250 - accuracy: 1.0000
43/50 [=====================>........] - ETA: 2s - loss: 0.0249 - accuracy: 1.0000
44/50 [======================>.......] - ETA: 1s - loss: 0.0248 - accuracy: 1.0000
45/50 [======================>.......] - ETA: 1s - loss: 0.0247 - accuracy: 1.0000
46/50 [======================>.......] - ETA: 1s - loss: 0.0245 - accuracy: 1.0000
47/50 [=======================>......] - ETA: 0s - loss: 0.0244 - accuracy: 1.0000
48/50 [=======================>......] - ETA: 0s - loss: 0.0241 - accuracy: 1.0000
49/50 [========================>.....] - ETA: 0s - loss: 0.0240 - accuracy: 1.0000
50/50 [========================>.....] - ETA: 0s - loss: 0.0240 - accuracy: 1.0000
50/50 [==============================] - 17s 335ms/step - loss: 0.0240 - accuracy: 1.0000 - val_loss: 0.1498 - val_accuracy: 0.9261
```

Рис. 3. Результат обучения сверточной нейронной сети на большом числе примеров

Далее мы несколько изменили исходные данные: вместо абстрактов с описанием месторождений железа были выбраны абстракты с общим описанием геологических массивов. Таких набралось 740. В этом случае мы выполнили бинарную классификацию – месторождения золота – геологические описания, не содержащие месторождений. Нам также успешно удалось выполнить тренировку этой модели, с такими же показателями точности. Но на этапе тестирования модели нас ждало разочарование. При выполнении классификации текстов с описанием объектов, на которых она обучалась (месторождения золота и описания геологических массивов), модель, как и в предыдущем случае, уверенно выполняла классификацию текстов. Но при попытке классификации текстов с описанием месторождений железа, которых не было в этой обучающей выборке, модель все их распознавала как описание месторождений золота. Это так называемая проблема обучения с нулевым результатом (zero-shot) [35], т. е. мы предложили язы-

ковой модели классифицировать объект, который она не видела на этапе обучения. Поэтому требуются дополнительное дообучение и корректировка модели или использование другого класса моделей (zero-shot learning), чтобы корректно обрабатывать такие ситуации. Это задача на будущее.

Задача выделения ключевых слов

Следующая задача в области обработки естественного языка, которую мы попытались решить, – это выделение ключевых слов из текстов геологической тематики. Извлечение ключевых слов (фраз) является высокоуровневым реферированием, позволяющим сжать большой документ до уровня емких коротких определений. В приведенном выше обзоре литературы, по крайней мере, две статьи решают эту задачу в области наук о Земле различными методами. Кроме того, мы можем отослать читателя к более подробным общим статьям по данной теме [36, 37].

Для решения этой задачи мы решили воспользоваться русскоязычной моделью T5 [38] и алгоритмом тренировки модели [39]. Эта модель относится к классу моделей генерирующего реферирования. Она распознает входной текст и создает новый текст на основе материала, на котором она была обучена. С ее помощью можно выполнять перевод текста, перефразирование, заполнение пропусков, восстановление, упрощение, ответы на вопросы по тексту, генерацию заголовков.

Как всегда, при обучении новой модели первым делом необходимо подготовить обучающую выборку. Мы воспользовались нашим архивом публикаций (https://repository.geologyscience.ru) и выбрали абстракты статей, которые сопровождаются ключевыми словами. Всего удалось получить 9320 записей.

Имеются две русскоязычные модели T5 – base и large. Как было отмечено [39], модель base не всегда корректно производит выделение ключевых слов, поэтому мы не стали экспериментировать с этой моделью, а сразу выбрали large модель. К сожалению, из-за размеров модели обучать ее на локальном компьютере оказалось проблематичным – не хватает ресурсов. Поэтому мы выбрали облачный сервис Яндекса – DataSphere [40]. Этот сервис обладает масштабируемой архитектурой и гибкой тарифной политикой. Мы использовали конфигурацию g2.4

(112 vCPU, 4 GPU A100). Обучение длилось около 20 минут.

Результаты тестирования модели

Абстракт:

«Изучены космоструктуры заангарский части Енисейского кряжа по материалам мультиспектральных космических систем Modis и Landsat ETM+. Выделены четыре системы кольцевых структур первого порядка, интерпретируемые как глубинные очаги гранитизации. Показаны закономерности размещения золотого оруденения в космогеологических структурах. Выделенные разноранговые космогеологические структуры находят отражение в аномальных структурах геохимических полей»

Реальные ключевые слова: золоторудные объекты; гранитизация; линейные структуры; кольцевые структуры; енисейский кряж;

Предложенные моделью: золото; енисейский кряж; енисейский кряж

Абстракт:

«Рассмотрены минералого-геохимические особенности каолинитовых прослоев (тонштейнов) Азейского месторождения Иркутского бассейна. Выявлена редкометалльная геохимическая специализация тонштейнов месторождения на РЗЭ, Y, Zr, Hf, U, Th, Ta, Sn, Ga, Cu, Pb, Se и Te. Проведен комплексный анализ возможных источников накопления первичного вещества тонштейнов. Приведено обоснование аэрогенного механизма накопления первичного материала каолинитовых прослоев. Обоснована вулканогенная модель их образования за счет пирокластического вещества кислого (липаритового) состава»

Реальные ключевые слова: иркутский угольный бассейн; азейское месторождение; минералогия; геохимия; уголь; тонштейны;

Предложенные моделью: иркутский бассейн; каолинитовые прослои;

Абстракт:

«Проведен литологический анализ верхнеюрско-нижнемеловой черносланцевой баженовской свиты двух районов северной части Хантейской гемиантеклизы и Межовского мегамыса. Первый из них расположен в пределах Пурпейско-Васюганского фациального района и отвечает глубоководной части палеобассейна, второй находится в Сильгинском фациальном районе и значи-

тельно более приближен к береговой линии. Разрезы баженовской свиты в районе Хантейской гемиантеклизы отличаются повышенной карбонатностью и кремнистостью, а также широким развитием аутигенного барита, что связывается с более интенсивным развитием фауны в этом районе. Баженовская свита на Межовском мегамысе характеризуется относительно низким содержанием карбонатных минералов и повышенным содержанием глинистого материала в породах»

Реальные ключевые слова: литология; верхняя юра; меловая;

Предложенные моделью: юрская; межовский мегамыс; баженовская свита;

Из приведенного примера видно, что обученная модель в целом неплохо справилась с выделением ключевых слов – нет явных промахов с неактуальными ключевыми словами. Но и недостатки налицо, в частности, это повторение ключевых слов и меньшее их количество. Главное в приведенном тесте, что все ключевые слова, предложенные моделью, хорошо соотносятся с текстом.

ЗАКЛЮЧЕНИЕ

Важный первый шаг в применении методов искусственного интеллекта для обработки текстов – это получение обучающей коллекции текстов. Во многих работах, связанных с обработкой текстов на русском языке, отмечаются проблемы с получением таких коллекций, особенно коллекций тематических. Поэтому еще раз отметим важность и своевременность создания ними архива публикаций с тематикой «Науки о Земле» (https://repository.geologyscience.ru/) [27], который позволил получить такие коллекции.

Бурное развитие в последние годы методов искусственного интеллекта, связанное с обработкой и генерацией текстов, открыло замечательные возможности по извлечению новых знаний из потока научной информации, которую очень трудно, а зачастую и невозможно переработать традиционным методом чтения. Приведенный краткий обзор литературы показывает, что применение этих методов в области наук о Земле находится в начальной фазе. Необходимо ознакомиться с данными методами, попробовать на простых задачах, выяснить области применения, их достоинства и недостатки. Наш первый опыт показал, что

методы обработки естественного языка действительно работают в области наук о Земле: можно получать непротиворечивые результаты. Необходимо дальнейшее изучение этих методов, чтобы можно было решать действительно серьезные задачи.

Работы выполняются в рамках Государственного задания ГГМ РАН по теме № 0140-2019-0005 «Разработка информационной среды интеграции данных естественнонаучных музеев и сервисов их обработки для наук о Земле», а также темы государственного задания № 1021061009468-8-1.5.1 «Цифровая платформа интеграции и анализа геологических и музейных данных».

СПИСОК ЛИТЕРАТУРЫ

1. *Kate A.* Машинное обучение и искусственный интеллект в геологии // Золотодобыча, №257, Апрель, 2020, пер.

2. *Kaplan A., Haenlein M.* Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence // Business Horizons. 2019. V. 62, No. 1. P. 15–25.

3. PROSPECTOR // URL: http://www.computing.surrey.ac.uk/ai/PROFILE /prospector.html (дата обращения 18.09.2023)

4. ESRI // URL: https://www.esri.com/en-us/home (дата обращения 18.09.2023)

5. USGS // URL: https://www.usgs.gov/ (дата обращения 18.09.2023)

6. *Родионов С.М., Сыркин В.К.* Экспертная прогнозирующая система «Олово» // Тихоокеанская геология. 1995. Т. 14, №5. С. 63–71.

URL: http://itig.as.khb.ru/POG/archive/1995/N5_1995.pdf

7. SOLSA Expert System // URL: https://solsa-dem-up.eu/en (дата обращения 17.09.2023)

8. GoldSpot // URL: https://www.alsglobal.com/en/consulting-and-analytics (дата обращения 18.09.2023)

9. SRK Consulting // URL: https://www.srk.com/ru/ (дата обращения 18.09.2023)

10. Maptek // URL: https://www.maptek.com/ (дата обращения 18.09.2023)

11. IOS Services Geoscientifiques // URL: https://www.iosgeo.com/en/ (дата обращения 18.09.2023)

12. Orefox // URL: https://orefox.com/ (дата обращения 18.09.2023)

13. Geolearn // URL: https://www.geolearn.ai/ (дата обращения 18.09.2023)

14. Datarock // URL: https://datarock.com.au/platform/ (дата обращения 18.09.2023)

15. *Baraboshkin E.E., Ismailova L.S., Orlov D.M., Zhukovskaya E.A., Kalmykov G.A., Khotylev O.V., Baraboshkin E.Yu., Koroteev D.A.* Deep Convolutions for In-Depth Automated Rock Typing // Computers & Geosciences. 2020. V. 135. https://doi.org/10.1016/j.cageo.2019.104330

16. *Nesteruk S., Agafonova J., Pavlov I., Gerasimov M., Latyshev N., Dimitrov D., Kuznetsov A., Kadurin A., Plechov P.* MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description // Computers & Geosciences. 2023. V. 178. https://doi.org/10.1016/j.cageo.2019.104330

17. Обработка естественного языка // URL: https://ru.wikipedia.org/wiki/Обработка_естественного_языка (дата обращения 18.09.2023)

18. *Jurafsky D., Martin J.H.* N-gram Language Models // Speech and Language Processing 3rd. 2021.

19. *Deng C., Zhang T., He Z., Chen Q., Shi Y., Zhou L., Fu L., Zhang W., Wang X., Zhou C., Lin Z., He J.* Learning Foundation Language Models for Geoscience Knowledge Understanding and Utilization // arXiv:2306.05064, 2023. URL: https://arxiv.org/abs/2306.05064v1

20. K2 model // URL: https://github.com/davendw49/k2?ysclid=lmswxywt6i750905070 (дата обращения 18.09.2023)

21. *Lawley C.J.M., Raimondo S., Chen T., Brin L., Zakharov A., Kur D., Hui J., Newton G., Burgoyne S.L., Marquis G.* Geoscience language models and their intrinsic evaluation // Applied Computing and Geosciences. 2022. V. 14, 100084. P. 1–10.

22. *Wang B., Ma K., Wu L., Qiu Q., Xie Z., Tao L.* Visual analytics and information extraction of geological content for text-based mineral exploration reports // Ore Geology Reviews. 2022. V. 144, 104818. P. 1–12.

23. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. V. 5. P. 177–187.

24. *Lawley C.J.M., Gadd M.G., Parsa M., Lederer G.W., Graham G.E., Ford A.* Applications of Natural Language Processing to Geoscience Text Data and Prospectivity Modeling // Natural Resources Research. 2023. V. 32, No. 4. P. 1503–1527.

25. *Fuentes I., Padarian J., Iwanaga T., Vervoort R.W.* 3D lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. V. 141, 104516.

26. *Qiu Qinjun, Xie Zhong, Wu Liang, Li Wenjia.* Geoscience keyphrase extraction algorithm using enhanced word embedding // Expert Systems with Applications. 2019. V. 125. P. 157–169.

27. *Патук М.И., Наумова В.В., Ерёменко В.С.* Цифровой репозиторий "geologyscience.ru": открытый доступ к научным публикациям по геологии России. // Электронные библиотеки. 2020. Т. 23, № 6. С. 1324–1338. https://doi.org/10.26907/1562-5419-2020-23-6-1324-1338

28. *Патук М.И., Наумова В.В.* Построение цифровой системы управления геологическими знаниями для поддержки научных исследований. // Электронные библиотеки. 2022. Т. 25, № 2. С. 148–158. https://doi.org/10.26907/1562-5419-2022-25-2-148-158

29. *Bourke D.* 08. Natural Language Processing with TensorFlow. URL: https://dev.mrdbourke.com/tensorflow-deep-learning/08_introduction_to_nlp_in_tensorflow/ (дата обращения 18.09.2023)

30. mrdbourke / tensorflow-deep-learning. URL: https://github.com/mrdbourke/tensorflow-deep-learning (дата обращения 18.09.2023)

31. Pdfreader 0.1.12. URL: https://pypi.org/project/pdfreader/ (дата обращения 18.09.2023)

32. spaCy URL: https://spacy.io/models/ru (дата обращения 18.09.2023)

33. Spacy-stanza. URL: https://spacy.io/universe/project/spacy-stanza (дата обращения 18.09.2023)

34. SberDevice. Как мы анализируем предпочтения пользователей виртуальных ассистентов Салют. URL: https://habr.com/ru/companies/ sberdevices/articles/547568/ (дата обращения 18.09.2023)

35. Zero-shot learning.
URL: https://en.wikipedia.org/wiki/Zero-shot_learning (дата обращения 18.09.2023)

36. *Ванюшкин А.С., Гращенко Л.А.* Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. С. 85–93.

37. *Pay T., Lucci F., Cox J.L.* An Ensemble of Automatic Keyword Extractors: TextRank, RAKE and TAKE // Computación y Sistemas. 2019. V. 23, No. 3. P. 703–710. https://doi.org/10.13053/CyS-23-3-3234

38. *Дале Д.* Многозадачная модель T5 для русского языка.
URL: https://habr.com/ru/articles/581932/ (дата обращения 18.09.2023)

39. *Данил*, keyT5 или генерация ключевых слов из текста.
URL: https://habr.com/ru/articles/599715/ (дата обращения 18.09.2023)

40. Yandex DataSphere. URL: https://datasphere.yandex.ru/?yc-skip-auth=1 (дата обращения 18.09.2023)

ARTIFICIAL INTELLIGENCE METHODS FOR SCIENTIFIC RESEARCH IN GEOLOGY

Mikhail I. Patuk[1] [0000-0003-3036-2275], Vera V. Naumova[2] [0000-0002-3001-1638]

[1, 2]*State Geological Museum named after Vladimir Vernadsky of RAS, Moscow*

[1]*patuk@mail.ru;* [2]*Naumova_new@mail.ru*

Abstract

A brief overview of some methods of artificial intelligence in the field of Earth sciences is given. The prospects of using these methods to obtain new knowledge are noted. The results of the authors' first attempts to apply natural language processing methods for processing scientific articles on geology are presented. The possibilities of developing work in this direction are discussed.

Keywords*:* Artificial intelligence, machine learning, natural language processing, geology.

REFERENCES

1. *Caté A.* Machine Learning and Artificial Intelligence for Mining Geoscience // Geological Association of Canada, 2019.

2. *Kaplan A., Haenlein M.* Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence // Business Horizons. 2019. V. 62. No. 1. P. 15–25.

3. PROSPECTOR // URL: http://www.computing.surrey.ac.uk/ai/PROFILE /prospector.html (дата обращения 18.09.2023) (date of access 18.09.2023)

4. ESRI // URL: https://www.esri.com/en-us/home (date of access 18.09.2023)

5. USGS // URL: https://www.usgs.gov/ (date of access 18.09.2023)

6. *Rodionov S.M., Syrkin V.K.* Expert forecasting system "Olovo" // Geology of the Pasofic Ocean. 1995. V. 14, No. 5. P. 63–71.
URL: http://itig.as.khb.ru/POG/archive/1995/N5_1995.pdf

7. SOLSA Expert System // URL: https://solsa-dem-up.eu/en (date of access 18.09.2023)

8. GoldSpot // URL: https://www.alsglobal.com/en/consulting-and-analytics (date of access 18.09.2023)

9. SRK Consulting // URL: https://www.srk.com/ (date of access 18.09.2023)

10. Maptek // URL: https://www.maptek.com/ (date of access 18.09.2023)

11. IOS Services Geoscientifiques // URL: https://www.iosgeo.com/en/ (date of access 18.09.2023)

12. Orefox // URL: https://orefox.com/ (date of access 18.09.2023)

13. Geolearn // URL: https://www.geolearn.ai/ (date of access 18.09.2023)

14. Datarock // URL: https://datarock.com.au/platform/ (date of access 18.09.2023)

15. *Baraboshkin E.E., Ismailova L.S., Orlov D.M., Zhukovskaya E.A., Kalmykov G.A., Khotylev O.V., Baraboshkin E.Yu., Koroteev D.A.* Deep Convolutions for In-Depth Automated Rock Typing // Computers & Geosciences. 2020. V. 135. https://doi.org/10.1016/j.cageo.2019.104330

16. *Nesteruk S., Agafonova J., Pavlov I., Gerasimov M., Latyshev N., Dimitrov D., Kuznetsov A., Kadurin A., Plechov P.* MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description // Computers & Geosciences. 2023. V. 178. https://doi.org/10.1016/j.cageo.2019.104330

17. Natural language processing URL: https://en.wikipedia.org/wiki/ Natural_language_processing (date of access 18.09.2023)

18. *Jurafsky D., Martin J.H.* N-gram Language Models // Speech and Language Processing 3rd. 2021.

19. *Deng C., Zhang T., He Z., Chen Q., Shi Y., Zhou L., Fu L., Zhang W., Wang X., Zhou C., Lin Z., He J.* Learning Foundation Language Models for Geoscience Knowledge Understanding and Utilization // arXiv:2306.05064, 2023. URL: https://arxiv.org/abs/2306.05064v1

20. K2 model // URL: https://github.com/davendw49/k2?ysclid=lmswxywt6i750905070 (date of access 18.09.2023)

21. *Lawley C.J.M., Raimondo S., Chen T., Brin L., Zakharov A., Kur D., Hui J., Newton G., Burgoyne S.L., Marquis G.* Geoscience language models and their intrinsic evaluation // Applied Computing and Geosciences. 2022. V. 14, 100084. P. 1–10.

22. *Wang B., Ma K., Wu L., Qiu Q., Xie Z., Tao L.* Visual analytics and information extraction of geological content for text-based mineral exploration reports // Ore Geology Reviews. 2022. V. 144, 104818. P. 1–12.

23. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. V. 5. P. 177–187.

24. *Lawley C.J.M., Gadd M.G., Parsa M., Lederer G.W., Graham G.E., Ford A.* Applications of Natural Language Processing to Geoscience Text Data and Prospectivity Modeling // Natural Resources Research. 2023. V. 32, No. 4. P. 1503–1527.

25. *Fuentes I., Padarian J., Iwanaga T., Vervoort R.W.* 3D lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. V. 141, 104516.

26. *Qiu Qinjun, Xie Zhong, Wu Liang, Li Wenjia.* Geoscience keyphrase extraction algorithm using enhanced word embedding // Expert Systems with Applications. 2019. V. 125. P. 157–169.

27. *Patuk M.I., Naumova V.V., Eryomenko V.S.* Digital repository "geology-science.ru": open access to scientific publications on russian geology // Russian Digital Library Journal. 2020. V. 23, No. 6. P. 1324–1338. https://doi.org/10.26907/1562-5419-2020-23-6-1324-1338

28. *Patuk M.I., Naumova V.V.* Building a digital geological knowledge management system to support scientific research // Russian Digital Library Journal. 2022. V. 25, No. 2. P. 148–158. https://doi.org/10.26907/1562-5419-2022-25-2-148-158

29. *Bourke D.* 08. Natural Language Processing with TensorFlow, URL: https://dev.mrdbourke.com/tensorflow-deep-learning/08_introduction_to_nlp_in_tensorflow/ (date of access 18.09.2023)

30. mrdbourke / tensorflow-deep-learning URL: https://github.com/mrdbourke/ tensorflow-deep-learning (date of access 18.09.2023)

31. Pdfreader 0.1.12. URL: https://pypi.org/project/pdfreader/ (date of access 18.09.2023)

32. spaCy. URL: https://spacy.io/models/ru (date of access 18.09.2023)

33. Spacy-stanza. URL: https://spacy.io/universe/project/spacy-stanza (date of access 18.09.2023)

34. SberDevice, How do we analyze the preferences of users of virtual assistants Salute. URL: https://habr.com/ru/companies/sberdevices/articles/547568/ (date of access 18.09.2023)

35. Zero-shot learning. URL: https://en.wikipedia.org/wiki/Zero-shot_learning (date of access 18.09.2023)

36. *Vanushkin A.S., Graschenko L.A.* Methods and algorithms for keyword extraction // New information technologies in automated systems. 2016. P. 85–93.

37. *Pay T., Lucci F., Cox J.L.* An Ensemble of Automatic Keyword Extractors: TextRank, RAKE and TAKE // Computación y Sistemas. 2019. V. 23, No. 3. P. 703–710. https://doi.org/10.13053/CyS-23-3-3234

38. *Dale D.* Multitasking model T5 for Russian. URL: https://habr.com/ru/articles/581932/ (date of access 18.09.2023)

39. *Danil*, keyT5 or generating keywords from text. URL: https://habr.com/ru/ articles/599715/ (date of access 18.09.2023)

40. Yandex DataSphere URL: https://datasphere.yandex.ru/?yc-skip-auth=1 (date of access 18.09.2023)

СВЕДЕНИЯ ОБ АВТОРАХ

ПАТУК Михаил Иванович – к. г.-м. н., и. о. н. с., научный отдел Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Michail I. PATUK – PhD, scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: patuk@mail.ru

ORCID: 0000-0003-3036-2275

НАУМОВА Вера Викторовна – д. г.-м. н., г. н. с., зав. Научным отделом Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Vera V. NAUMOVA – Prof., head of SGM scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: naumova_new@mail.ru

ORCID: 0000-0002-3001-1638

Материал поступил в редакцию 22 сентября 2023 года