УДК 004.65(075.8)+004.89

ЗАПРОСЫ К НЕРЕЛЯЦИОННЫМ ДАННЫМ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ НА ОСНОВЕ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ

А. О. Еркимбаев[1] [0000-0002-5239-2208], В. Ю. Зицерман[2] [0000-0003-3327-3139], Г. А. Кобзев[3] [0000-0001-9987-1823]

[1-3]*Объединенный институт высоких температур РАН, г. Москва, Россия*

[1]adilbek@jiht.ru, [2]vz1941@mail.ru, [3]gkbz@mail.ru

Аннотация

В работе рассмотрены новые возможности организации запросов на естественном языке к научным локальным базам данных нереляционного типа. Проведенный анализ исследований, выполненных за последние годы, показал активное внедрение запросов на естественном языке к базам данных различного типа. Отмечено активное применение методов машинного обучения (нейронных алгоритмов). Показано широкое использование в последние два года большой языковой модели для подготовки запросов в различных языковых средах и областях знаний. Проведено исследование новых возможностей графовой базы данных AllegroGraph по использованию больших языковых моделей для организации поиска на естественном языке. Функционал базы данных изучен на примере системы метаданных по теплофизическим свойствам веществ в форме предметной онтологии «Термаль». Тестирование поисковых запросов в двуязычной (английская и русская) среде базы данных выявило в целом преодолимые проблемы и дает хорошие надежды на дальнейшее применение новых прикладных сервисов с использованием больших языковых моделей.

Ключевые слова*: запрос на естественном языке, большая языковая модель, эмбеддинг, нереляционные базы данных, графовая база данных, онтология предметной области.*

ВВЕДЕНИЕ

Работа посвящена расширению функциональности баз данных (БД), работающих при поддержке средств искусственного интеллекта (ИИ). Эта проблема изучается последние годы с преимущественной ориентацией на реляционные (SQL) БД с их хорошо структурированной системой хранения и поиска. Наш опыт по систематизации естественно-научных данных показал существование потребности в полуструктурированных данных с изменчивой структурой, что требует перехода на нереляционные (NoSQL) БД [1, 2].

Активное внедрение средств ИИ открывает новые возможности при работе с подобного рода данными, в частности в организации поиска. Одной из них является возможность организации запросов на естественном языке (ЕЯ), что избавляет пользователя от необходимости точного знания классификации и лексики предметной области, а также сложных правил составления поисковых запросов для NoSQL БД.

Применение запросов на ЕЯ к хранилищам данных имеет давнюю историю и восходит к информационной системе, часто упоминаемой как LUNAR [3], созданной в 1972 г. Полное наименование системы “The Lunar Sciences Natural Language Information System”. Она была предназначена для обычных пользователей и отвечала на вопросы о химическом анализе лунных пород, полученных с «Аполлона-11». Система состояла из трех основных компонентов: формальной грамматики общего назначения, синтаксического анализатора для большого подмножества естественного английского языка и компоненты семантической интерпретации, управляемой правилами.

ОБЗОР СОВРЕМЕННЫХ ПОДХОДОВ

На примере ряда работ установлены общие принципы составления запросов на ЕЯ. Так, авторы [4] подчеркивают, что при решении этой задачи всегда требуются: а) преобразование предложений на ЕЯ в информацию, пригодную для машинной обработки; б) транслирование запросов к БД, составленных на основании информации, извлеченной из текста. Преобразование текстов на ЕЯ предложено выполнять методами компьютерной лингвистики в полуавтоматическом или автоматическом режимах с применением графематического, морфологического и синтаксического анализов [5]. Итог реализации данного шага –

синтаксическое дерево с определенным набором объектов и свойств, которое впоследствии может быть преобразовано в запросы на формальном языке. Классифицируя существующие БД по их типу как дореляционные, реляционные и постреляционные (NoSQL), авторы [4] предложили наиболее реализуемую и оптимальную организацию запросов на ЕЯ к реляционным и графовым БД (которые относятся к NoSQL). С учетом выделенных вариантов БД ими указаны два пути обработки полученного синтаксического дерева: а) с помощью семантических сетей, отражающих связи между объектами; б) с помощью средств логического программирования, например Prolog. В дальнейшем результаты [4] были реализованы в системе запросов на ограниченном ЕЯ к реляционным БД [6].

Особенность другой работы [7] состояла в том, что использовалась семантическая модель БД как на этапе формирования естественно-языкового интерфейса, так и в ходе его эксплуатации. Первоначальный процесс обработки естественно-языкового запроса пользователя состоял из последовательного выполнения анализа. Следующий шаг обработки запроса на ЕЯ заключался в построении его морфологического, синтаксического и семантического представлений. При этом семантическое представление естественно-языкового запроса пользователя строилось на основе семантической модели БД, имеющей обязательное текстовое представление, доступное для ручной правки экспертами или машинной обработки. На основе этого представления в дальнейшем формировался SQL-запрос к БД.

Отметим, что работы, названные выше, не использовали методы машинного обучения (например, нейронные алгоритмы) для формирования запросов и делали акцент на семантические модели и логическое программирование.

Запрос на ЕЯ и выделенные после его разбора объекты синтаксического и семантического анализа (слова, части речи и предложения) представляют собой категориальный (нечисловой) тип данных. Как правило, для дальнейшей их обработки компьютером требуется их векторизация – преобразование текста в числовой формат, который могут понимать и обрабатывать программы, например алгоритмы машинного обучения. Активное применение нейронных сетей привело к созданию методами машинного обучения процедур векторизации текста [8]. Эта процедура преобразования текста получила название текстового *эмбеддинга* (text embeddings), или «встраивания текста». При этом различают

векторизацию слова (word embedding) или предложений (sentence embeddings). Существуют эмбеддинги иных типов данных (например, изображений, графовых структур и т. д.). К наиболее распространенным сейчас видам текстовых эмбеддингов в машинном обучении относят Word2Vec [8], Glove [9] и BERT [10].

Развитие методов глубокого обучения в настоящее время привело к тому, что встраивание текста или эмбеддинг стало основополагающей технологией в области обработки ЕЯ, способствующей прогрессу в решении множества последующих задач, связанных с языком. При этом по-прежнему одной из важных задач является определение семантического сходства текстов, что требует векторизации (встраивания) текста для вычисления сходства. В таких технологиях важную роль играют размеры и качество текстовых массивов или корпусов при организации обучения. На их основе создано множество фреймворков с готовыми решениями для создания запросов данных. Так, выполненный в обзоре [11] анализ 35 фреймворков, разработанных в период с 2008 по 2018 г., учитывал поддержку языка, эвристические правила, функциональную совместимость, объем данных и оценку производительности. Оказалось, что 70% запросов на ЕЯ было выполнено для SQL, а на долю NoSQL приходилось только 15% (SPARQL), 10% (CYPHER) и 5% (GREMLIN).

Результаты [11] трудно обобщить, поскольку тестировались они на различных БД и в разных предметных областях. Но при этом следует подчеркнуть, что подавляющее большинство фреймворков, рассмотренных в [11], выполняет запросы на ЕЯ к данным на английском языке, исключение составил лишь один продукт, работающий с арабским языком.

В работе [12] представлен обзор инструментов (фреймворков, платформ) по преобразованию запросов на ЕЯ к формату SQL с применением нейронных сетей, в котором приведены результаты тестирования на единой реляционной БД SPIDER [13]. Было проведено тестирование девяти моделей формирования запросов на ЕЯ. Наиболее эффективной моделью оказалась RAT-SQL, объединенная с методом векторизации BERT [10], которая обеспечивала точность 65.6% при выполнении междоменных SQL-запросов. Остальные модели обеспечивали точность формирования запросов от 30% до 50%. Важно отметить, что все методы организации запросов на ЕЯ, рассмотренные в обзорах [11, 12], были также созданы для работы с английским языком.

ПРИМЕНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ЗАПРОСОВ ДАННЫХ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Рассмотренные выше источники не отражают тенденции последних 3–5 лет, связанные с применением больших языковых моделей (Large Language Model, LLM [14]) для выполнения запросов к БД на ЕЯ. По типу это модели глубокого обучения, которые обучены на огромных объемах данных и содержат более миллиарда параметров. Благодаря экстремальному числу параметров они реализуют возможности распознавать, переводить, прогнозировать или генерировать текст, как и другой контент (изображения, звук и т. д.). В 2022 г. появился продукт ChatGPT [15], разработанный компанией OpenAI на базе ИИ и основанный на LLM-модели. Эта система способна работать в диалоговом режиме, отвечать на вопросы и генерировать тексты, что особенно важно, на разных языках, включая русский, относящиеся к различным областям знаний. Важной особенностью ChatGPT является также возможность генерации по запросу программ на различных языках программирования. В течение всего нескольких лет многие разработчики программных продуктов, оперирующих различными типами БД, создали собственные сервисы для запросов на ЕЯ на основе LLM. Отметим следующие из них.

- Сервис Microsoft Copilot [16] для выполнения запросов на ЕЯ к БД SQL Azure на контролируемом внешнем ресурсе Microsoft с платными и бесплатными возможностями (доступен на территории РФ только через VPN). Запросы можно выполнять на более чем 10 языках, в том числе на русском. Сервис основан на применении платформы вычислительного кластера OpenAI.
- Функциональные возможности по созданию запросов на ЕЯ к NoSQL документной базе данных MongoDb [17]. Запросы на ЕЯ на версии MongoDb Compass доступны, начиная с версии 1.40.x. В качестве текущего поставщика запросов используется Azure OpenAI.
- GraphDB предлагает палитру моделей ИИ с набором аналитических возможностей и инструментов [18]. GraphDB предоставляет инструменты LLM, использующие спецификацию OpenAI API (Application Programming Interface).

- Функционал LLM-ориентированных сервисов по созданию запросов на ЕЯ в графовой БД AllegroGraph [19]. Сервисы основаны на применении платформы вычислительного кластера OpenAI API.
- Сервис Stardog Voicebox в графовой БД Stardog [20] – интеллектуальный помощник по знаниям, работающий на базе LLM и автономных агентов для предоставления основных сервисов. Использует свою кластерную среду LLM и работает только в облачном пространстве Stardog Cloud.

Таким образом, четыре из пяти перечисленных примеров программных продуктов при использовании запросов на ЕЯ основаны на обязательном применении возможностей OpenAI [15].

Для Объединенного института высоких температур (ОИВТ) РАН с его обширной системой БД по свойствам веществ и материалов представляют особый интерес последние достижения в организации запросов к нереляционным БД с характерной для них сложной структурой, отражающей специфику предметной области. Ниже приведены результаты начальных экспериментов по реализации запросов на ЕЯ к онтологическим моделям, размещенным на платформе графовой БД AllegroGraph, указанной ранее в списке новых предлагаемых сервисов.

В ОИВТ функционирует несколько БД по теплофизическим свойствам веществ в текстовом формате ISO 2709 [21] дореляционного типа, в частности БД «Термаль». Необходимость переноса подобных данных в Интернет с реализацией на новой программно-аппаратной технологии привела нас к разработке системы метаданных в виде онтологической модели «Термаль» с применением технологий семантического веба. В качестве платформы для онтологической модели используется нереляционная графовая БД AllegroGraph, а носителем основной части данных является нереляционная БД MongoDb [17].

С целью оценки возможного применения были проведены тестовые испытания новых сервисов, предлагаемых AllegroGraph, с LLM-технологиями формирования запросов к онтологии на ЕЯ. В качестве объекта испытаний был выбран один из текущих вариантов разрабатываемой онтологии «Термаль».

После загрузки онтологической модели на серверный вариант БД AllegroGraph в облаке и для применения сервиса LLM по работе с ЕЯ был приобретен ключ доступа к прикладным функциям API кластера OpenAI. Функционал LLM для создания запросов на ЕЯ позволил провести векторизацию онтологии

«Термаль» при ее сохранении в векторном хранилище БД AllegroGraph. В настоящее время БД AllegroGraph предлагает на своей платформе два новых сервиса для создания запросов на ЕЯ с применением в нереляционной среде запросов SPARQL.

1. Сервис так называемых «магических» предикатов и функций, которые можно использовать в запросах, связанных с LLM. Магический предикат может использоваться в позиции предиката в запросах SPARQL, а функции могут преобразовывать значение переменной, что значительно расширяет возможности запросов SPARQL. Предложен ряд «магических предикатов», таких, например, как:

- llm: response: как функция, так и предикат, в зависимости от того, хотим ли мы, чтобы LLM возвращала один элемент или список элементов;
- llm:askMyDocuments: предикат высокого уровня, позволяющий запрашивать у LLM информацию о локальном хранилище векторов;
- llm:node: функция для генерации уникального URI для текстового литерала. llm:NearestNeighbor: предикат, который работает в хранилище векторов AllegroGraph. Он принимает в качестве входных данных строку или запрос и находит наилучшие совпадения в хранилище векторов;
- llm:askForTable: предикат высокого уровня, позволяющий запрашивать у LLM информацию и возвращать результаты в табличной форме.

Здесь конструкция llm: – это предопределенный префикс в системе AllegroGraph (http://franz.com/ns/allegrograph/8.0.0/llm/).

2. Сервис функции преобразования запроса на ЕЯ в запрос SPARQL – Natural Languge to SPARQL (новая функция, находящаяся еще в стадии разработки). Для использования сервиса необходимо создать специализированную векторную БД (VDB), в которой хранятся пары запросов на ЕЯ и соответствующие им SPARQL-запросы. База VDB напрямую связана с триплетным хранилищем AllegroGraph и действует как хранилище сопоставлений между тем, как пользователи могут задать вопрос на ЕЯ, и тем, как этот запрос должен быть выражен в SPARQL. По мере того как сохраняется все больше сопоставлений в VDB, сервис становится все более способным преобразовывать запросы пользователей в более точные запросы SPARQL.

В качестве примера общей работоспособности предлагаемых системой БД AllegroGraph возможностей выполнения запросов на русском («Список республик в России») и английском (“List the USA states”) ЕЯ на рис. 1 представлены результаты применения использования «магического» предиката llm:response из первого сервиса к глобальным ресурсам интернета через API функции OpenAI.

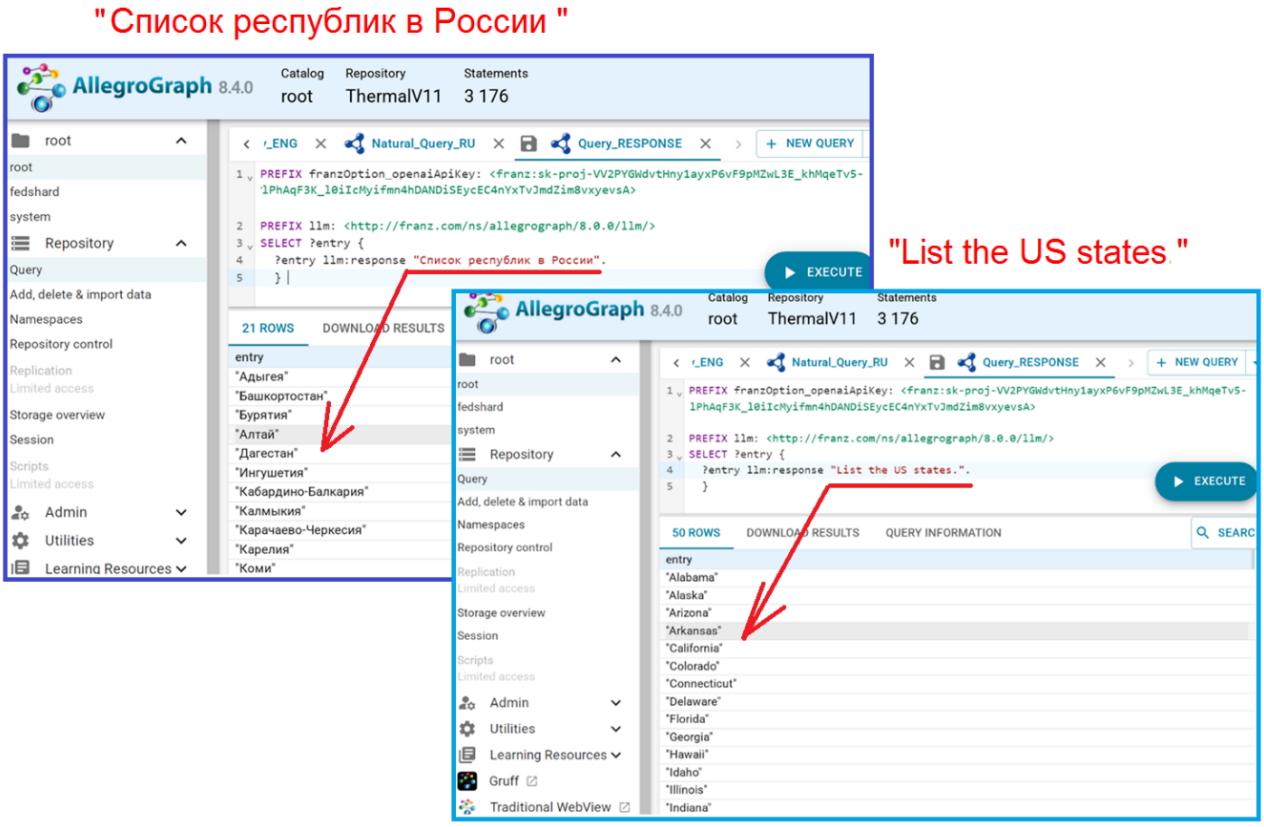

Рис. 1. Сканы интерфейсов БД AllegroGraph [19] при выполнении запросов на русском и английском ЕЯ к глобальным ресурсам интернета с использованием «магического» предиката llm:response.

Далее на рис. 2 представлены интерфейсы, демонстрирующие технологию реализации на платформе БД AllegroGraph двух разных сервисов для создания запросов на ЕЯ к локальным ресурсам через API функции OpenAI.

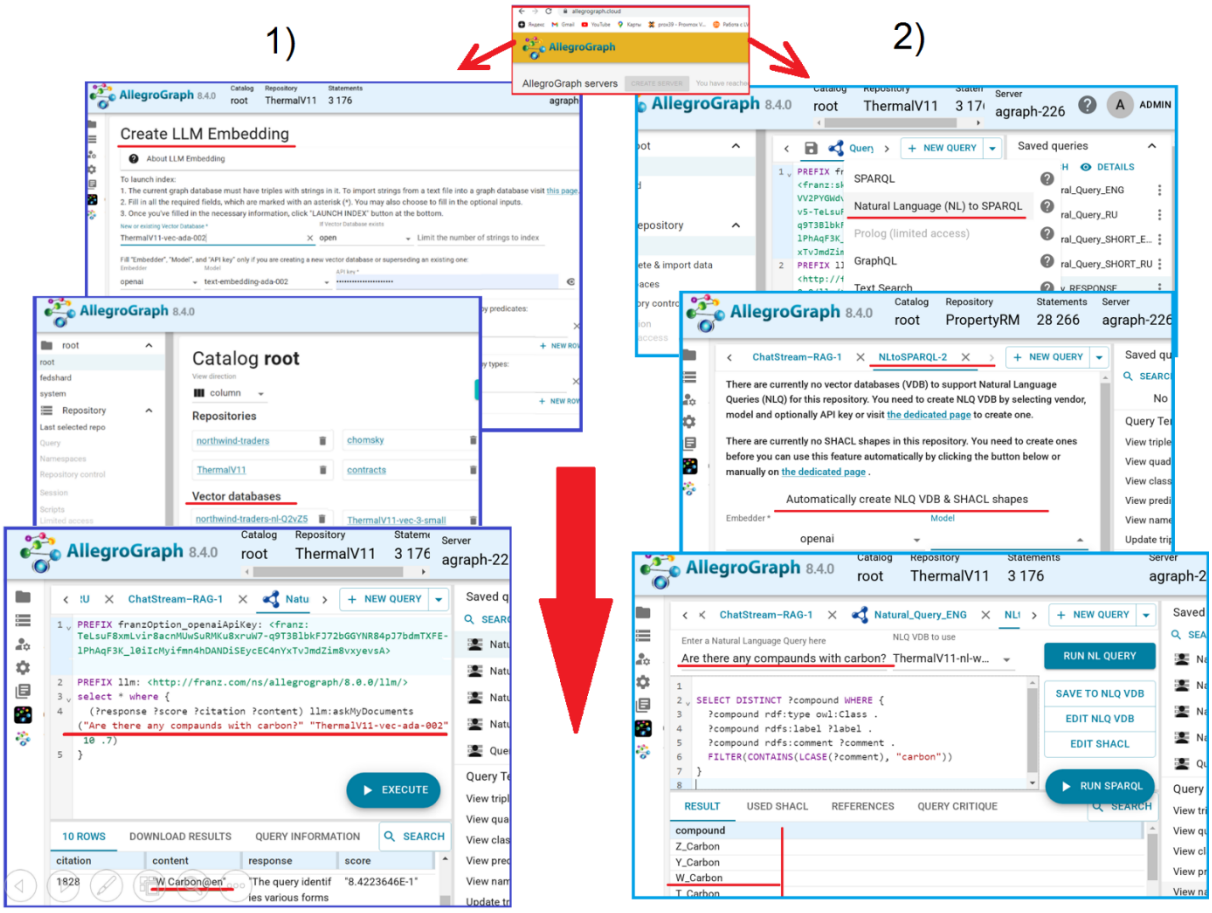

Рис. 2. Последовательности сканов интерфейсов БД AllegroGraph [19], реализующие два сервиса для создания запросов на ЕЯ к локальным ресурсам: 1) сервис «магических» предикатов; 2) сервис функции преобразования запроса на ЕЯ в SPARQL запрос.

Были выполнены тестовые запросы с использованием двух сервисов организации запросов на ЕЯ к локальной системе метаданных в форме онтологии «Термаль», загруженной в нереляционную БД AllegroGraph. В качестве проверочных данных были выбраны произвольные выражения на русском и английском языках для формирования запроса: «Есть ли соединения с углеродом?»; “Are there any compaunds with carbon?”.

Для проверки первого сервиса организации запросов на ЕЯ был использован «магический предикат» llm:askMyDocuments. В условиях запроса в атрибутах функции предписано выдать ближайшие 10 результатов с оценкой приближения не ниже 0.7 (1.0 – полное совпадение, 0.0 – отсутствие совпадения). На рис. 3 представлены интерфейсы запросов на ЕЯ с результатами запроса к векторному образу «ThermalV11-vec-ada-002» онтологии.

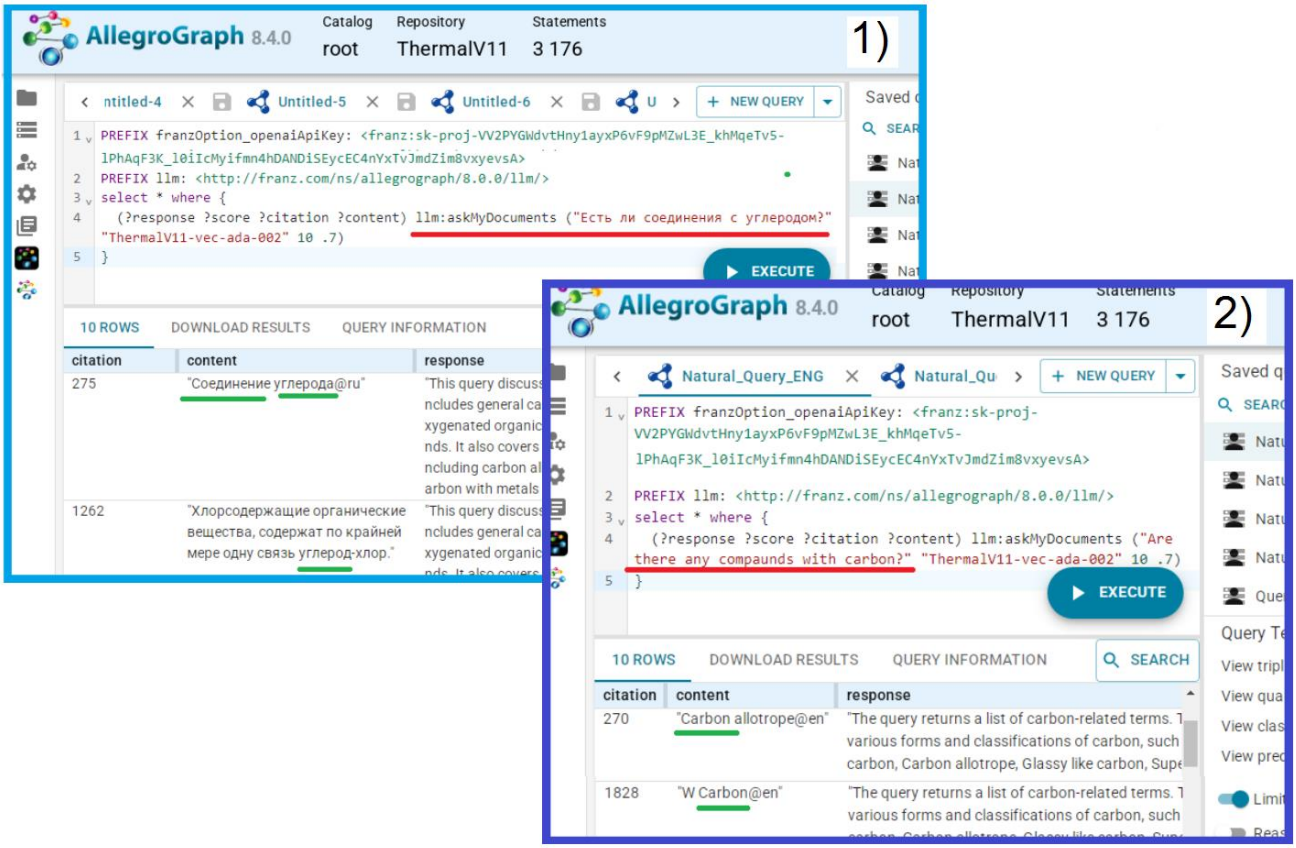

Рис. 3. Сканы интерфейса БД AllegroGraph с запросами на русском (1) и английском (2) языках к онтологии «Термаль»

На рис. 4 представлены четыре примера найденных классов онтологии «Термаль» с содержанием различных аннотационных свойств. На рис. 5 показаны (рядом, для сравнения) полные списки результатов поиска на русском и английском ЕЯ с указанием содержания найденного контента, оценки приближения результата применяемой функциональной модели LLM и соответствующего контенту класса онтологии. Для пояснения на рис. 3–5 зеленным цветом выделены совпавшие поисковые слова из запроса на ЕЯ.

Отметим следующие важные моменты в полученных результатах.

1. Поиск, выполненный в векторном образе онтологии, выдал 10 ближайших результатов, содержащих поисковые слова из запроса на ЕЯ с максимальной оценкой приближения 0.8941 для запроса на русском языке и 0.8531 на английском для классов “Carbon_ME” и “Elemental_Carbon_ME” соответственно по их аннотационным свойствам rdfs:label "Соединение углерода"@ru и

rdfs:label "Elemental carbon"@en (см. рис. 3 и 5). Это свидетельствует о формальной работоспособности функционала LLM графовой БД для выполнения запросов на русском и английском ЕЯ.

2. Из 10 результатов только в одном случае запросы на разных языках к онтологии показали одинаковый результат – класс “Carbon_allotrope” благодаря наличию аннотационного свойства на обоих языках: rdfs:label "Carbon allotrope"@en, rdfs:label "Аллотроп углерода"@ru (см. п. 2 на рис. 4 и строки с желтым фоном на рис. 5). Отмеченный результат показал, что не все классы тестовой онтологии обладают полным набором аннотационных свойств на русском и английском языках, что и объясняет наличие всего одного совпадения.

3. При формально правильном совпадении поисковых слов «соединения» и “compounds” со значениями аннотационных свойств класса “Oxygen_ME” (rdfs:label "Соединение кислорода"@ru см. п. 4 рис. 4) и класса “Addition_compound” (rdfs:label "Addition compound"@en см. рис. 5) результат оказался неверным по смыслу запроса. Это указывает на важность выбора выражений на ЕЯ при формировании запроса и более тщательной подготовки значений свойств онтологии. В то же время это является указателем необходимости проверки всех возможностей векторизации онтологии в данном функционале LLM графовой БД: отдельных слов, словосочетаний и предложений.

4. Результаты запроса на русском языке в большинстве своем получены на основе значений аннотационного свойства rdfs:comment @ru (комментарий) на русском ЕЯ и аннотационного свойства rdfs:label @en (ярлык) на английском ЕЯ, см. рис. 5. Это обстоятельство еще раз подчеркивает важность подготовки свойств и содержания онтологии на разных языках.

Полученные результаты показывают, что одинаковый по смыслу вопрос дает различный набор ответов на разных языках (см. рис. 5), что с формальной точки программирования не является нарушением. Однако это указывает на неполноту и неустойчивость подобной модели поиска данных на ЕЯ, особенно при работе с семантическим объектом – онтологией. Но отмеченные выше моменты при анализе результатов дают объяснение возникшей проблемы и пути к ее устранению. Полноценная подготовка значений свойств онтологий, выбор кор-

ректных выражений запросов на ЕЯ и использование всех возможностей векторизации должны обеспечить приемлемое совпадение ответов на различных языках и помочь преодолеть этот недостаток модели поиска.

В целях проверки были проведены частичная коррекция онтологии добавлением недостающих аннотационных свойств на английском и русском языках для классов, описывающих соединения углерода, и уточнение выражений при формулировке запроса. Использование пробных уточняющих выражений для запросов на русском и английском ЕЯ в этом случае приводит к значительному росту совпадающих ответов независимо от языка запроса. Таким образом, предварительные результаты проверки дают обнадеживающие перспективы преодоления обнаруженной проблемы рассмотренной модели поиска на ЕЯ в двуязычной предметной онтологии по теплофизическим свойствам.

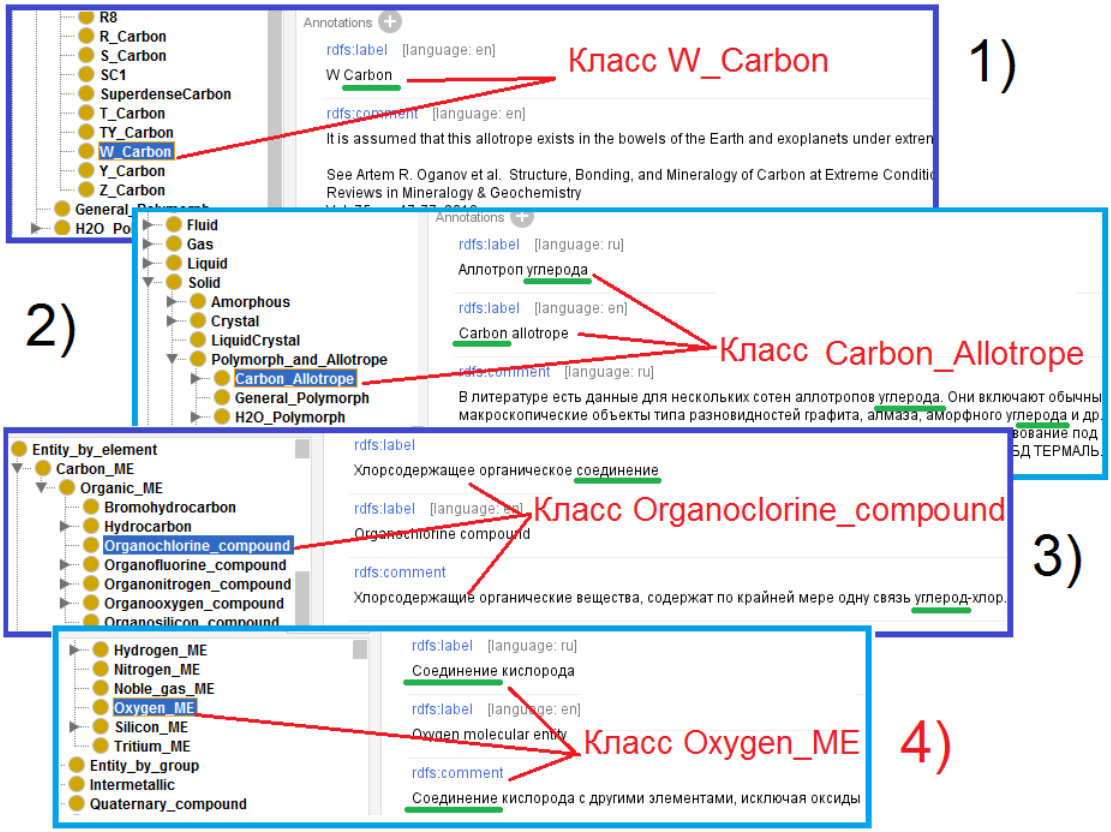

Рис. 4. Фрагменты онтологии «Термаль» с указанием найденных классов по запросу на английском и русском ЕЯ. Представлено содержание аннотационных свойств классов rdfs:label, rdfs:comment

Результаты запроса на русском языке "Есть ли соединения с углеродом"

Содержание контента, ЗЕЛЕННЫМ выделены найденные объекты	score	Имя класса в онтологии "Термаль"
Соединение углерода	0.8941	Carbon_ME
Хлорсодержащие органические вещества; содержат по крайней мере одну связь ***углерод***-хлор.	0.8671	Organochlorine_compound
Кислородсодержащие органические вещества; содержат по крайней мере одну ***углерод***-кислородную связь.	0.867	Organooxygen_compound
Фторсодержащие органические вещества; содержат по крайней мере одну связь ***углерод***-фтор.	0.859	Organofluorine_compound
Углеводород; содержащий замкнутую в кольцо цепь атомов ***углерода***.	0.8578	Cyclic_hydrocarbon
Соединение кислорода	0.8565	Oxygen_ME
Азотсодержащие органические вещества; содержат по крайней мере одну ***углерод***-азотную связь.	0.8544	Organonitrogen_compoun
Аллотроп ***углерода***	0.8536	Carbon_Allotrope
Карбиды; соединения ***углерода*** с металлами; а также с бором и кремнием.	0.8525	Carbide
3-элементное соединение переходного металла с ***углеродом*** и серой	0.8502	Carbosulphide

Результаты запроса на английском языке "Are there any compounds with carbon"

Содержание контента, ЗЕЛЕННЫМ выделены найденные объекты	score	Имя класса в онтологии "Термаль"
Elemental ***carbon***	0.8531	Elemental_carbon_ME
W ***Carbon***	0.8506	W_Carbon
Carbon allotrope	0.8495	Carbon_Allotrope
Amorphous ***Carbon***	0.8386	AmorphousCarbon
T ***Carbon***	0.8403	T_Carbon
R ***Carbon***	0.8394	R_Carbon
M-***carbon***	0.839	M_Carbon
Z ***Carbon***	0.8374	Z_Carbon
Addition ***compound***	0.8346	Addition_compound
Superdense ***carbon***	0.8336	SuperdenseCarbon

Рис. 5. Сравнение результатов запроса на русском и английском ЕЯ к онтологии «Термаль». На рисунке: желтый цвет фона – совпадение результатов поиска на русском и английском языках по классу в онтологии; красная линия подчеркивания – найденные классы не соответствуют смысловому содержанию запроса.

Результаты проверки второго сервиса организации запросов на ЕЯ представлены на сканах интерфейсов БД AllegroGraph, см. рис. 6–8. Они показали принципиальную работоспособность предлагаемого сервиса и возможности его самоулучшения в процессе применения. Проблема одинаковых по смыслу запросов на русском и английском языках проявилась в различном содержании составленных сервисом запросов SPARQL, и соответственно, в получаемых результатах. Так, на рис. 6 представлены два варианта ответа на запрос на русском ЕЯ, отражающие различные решения, предлагаемые сервисом. В одном случае сервис принимает решение использовать в составленном запросе SPARQL слово «углерод», а в другом – его перевод на английский язык “carbon”, что, разумеется, привело к разным результатам. Данная проблема была нами описана выше, при обсуждении результатов применения сервиса «магических» предикатов.

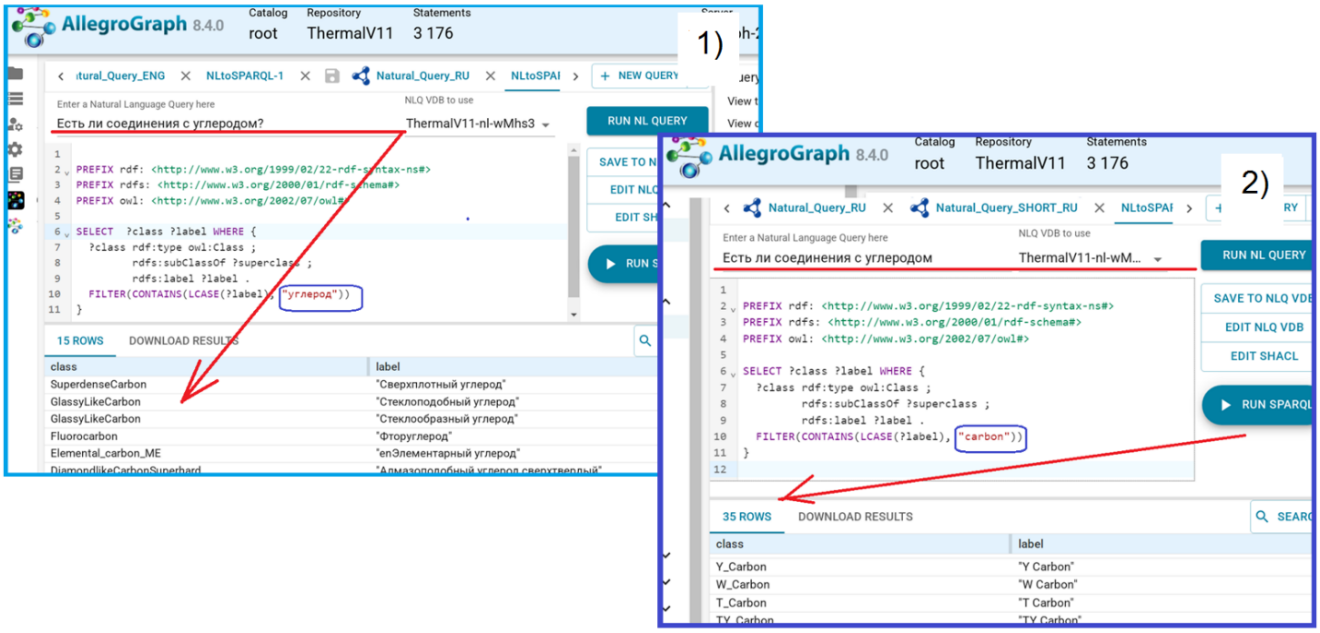

Рис. 6. Варианты запроса на ЕЯ на русском языке, выполненные разными запросами SPARQL. В запросе использовано слово: 1) «углерод»; 2) “carbon”

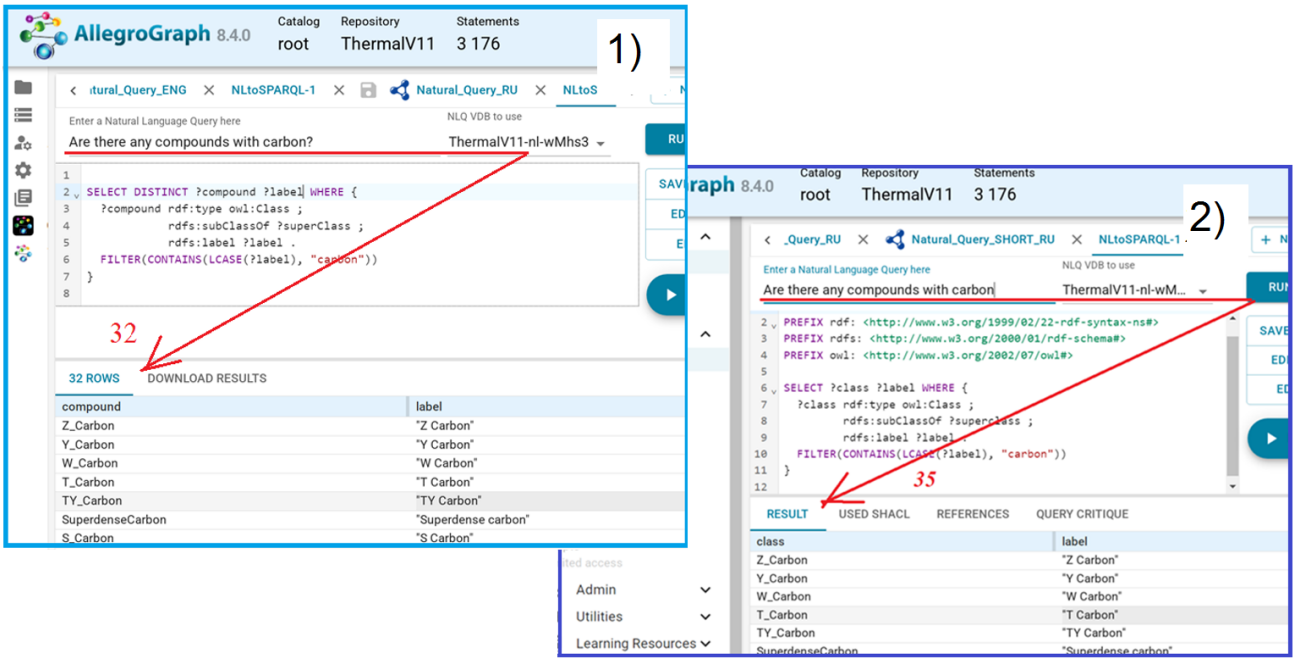

Рис. 7. Примеры уточнения запроса в процессе выполнения запроса на английском языке: 1) правильный результат 32 записи; 2) первоначальный запрос с неверным результатом в 35 записей.

На рис. 7 показан пример, когда в процессе многократного применения запросов произошли уточнение составленного запроса, его так называемое «самоулучшение» и выдача верного результата. Это выразилось в том, что во вновь сформулированном запросе SPARQL была добавлена опция уникальности (не повторяемости) “DISTINCT”, что привело к верному результату.

Наиболее интересный результат был отмечен при получении ответа на запрос ЕЯ количественного характера на двух языках, см. рис. 8. Запрос на английском ЕЯ выдал верный ответ – 32, в отличие от запроса на русском ЕЯ – 35. Данное обстоятельство в целом подтверждает особенность модели LLM, заключающуюся в более точных ответах на запросы на английском ЕЯ в глобальном интернете.

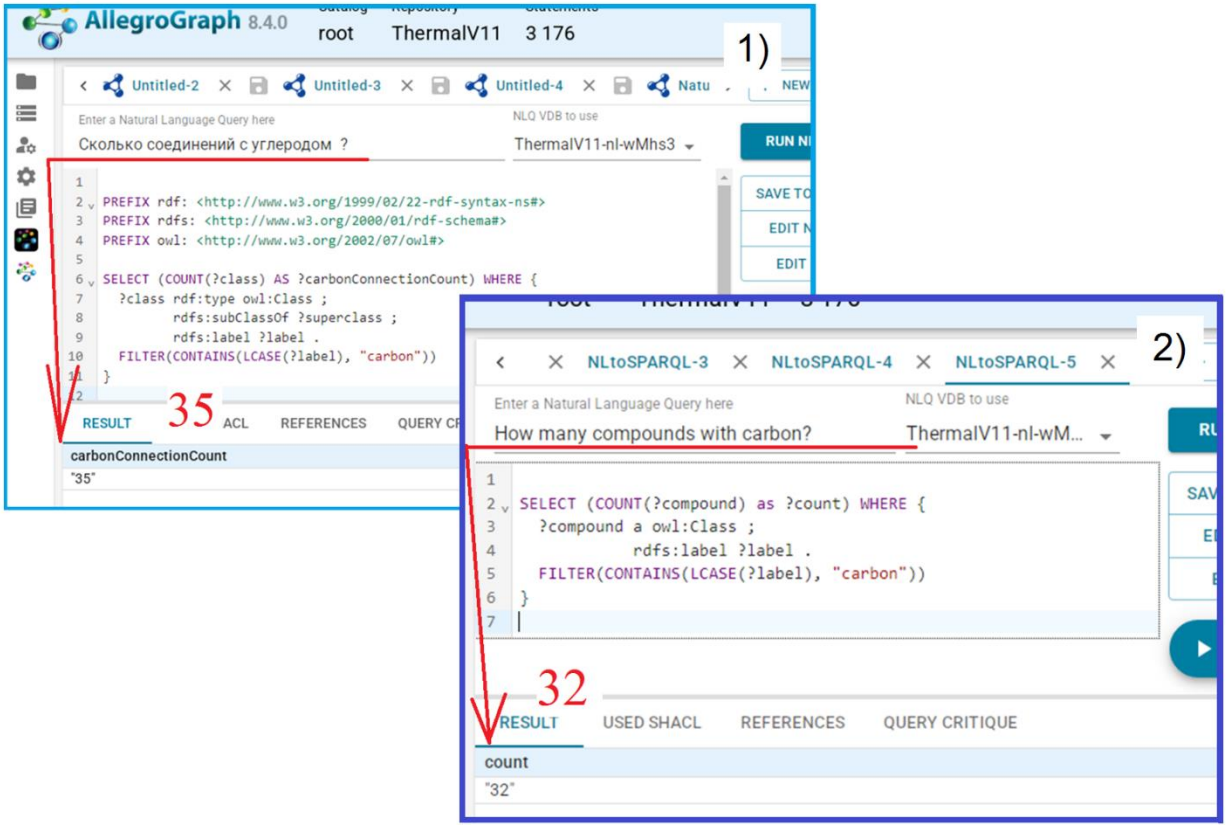

Рис. 8. Пример выполнения запроса на ЕЯ о количестве записей: 1) запрос на русском языке «Сколько соединений с углеродом?» – 35; 2) запрос на английском языке “How many compounds with carbon?” – 32.

Результаты, приведенные на рис. 7 и 8, объяснимы, если мы посмотрим на дерево онтологии «Термаль» (1 на рис. 9) и список из 35 классов (2 на рис. 9), полученные в запросе без применения опции “DISTINCT” из рис. 7 (2). Здесь на рис. 9 выделены три класса онтологии, имеющие дубли в силу того, что они имеют нескольких «родителей».

Так, например, класс «K6 Carbon» является одновременно подклассом класса «Аллотроп углерода» и подклассом класса «Металлический углерод».

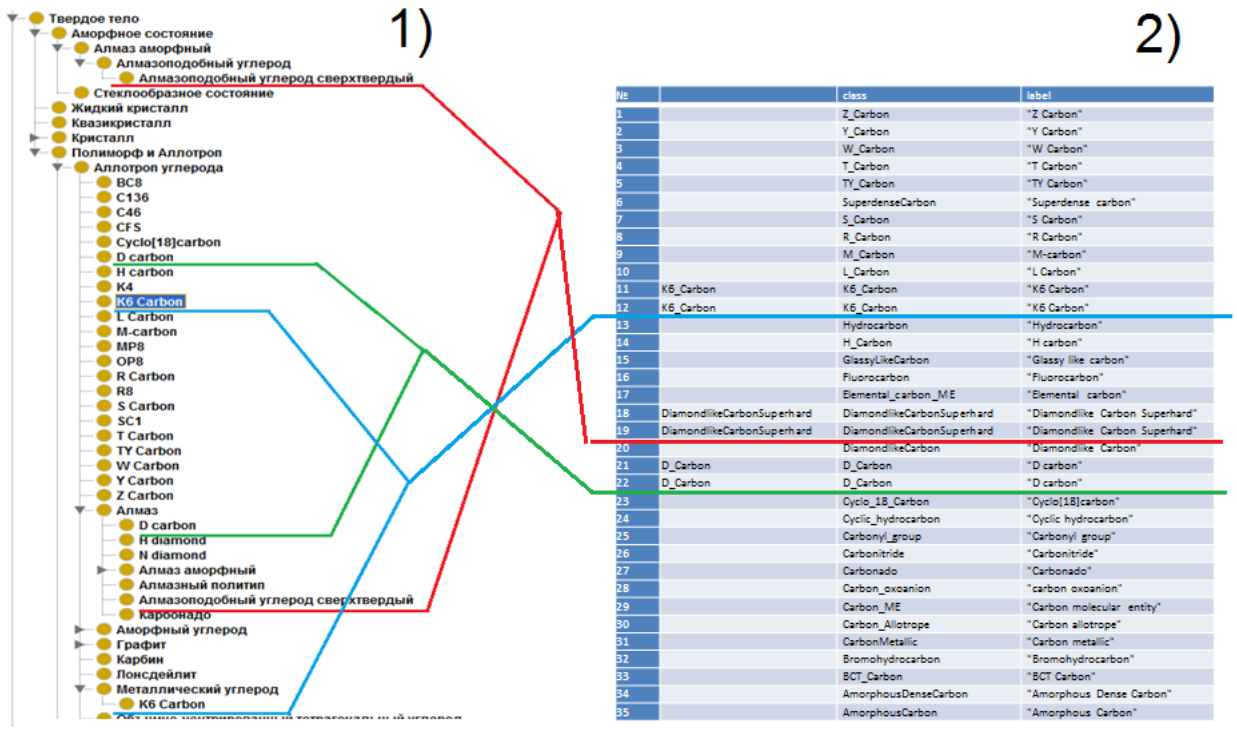

Рис. 9. Демонстрация фрагмента онтологии и списка результатов, полученных после выполнения запроса на ЕЯ, с примерами дублирования классов: 1) фрагмент онтологии; 2) список из 35 классов. На рисунке одинаковыми цветами отмечены продублированные классы в онтологии и списке результатов.

Таким образом, в целом тестирование показало работоспособность сервиса функции преобразования запроса на ЕЯ в запрос SPARQL.

ЗАКЛЮЧЕНИЕ

Проведенный анализ современного состояния средств и технологий, способных реализовать на ЕЯ запросы к БД различного типа, подтвердил развитие

возможностей в решении этих задач средствами LLM и их активное внедрение в прикладные сервисы во множество применяемых БД. Следует особо отметить одно из преимуществ LLM, которое состоит в способности работать на разных языках, включая русский. При этом языковое многообразие допустимо как в самом запросе, так и в семантическом ресурсе, которому направлен запрос. Другое преимущество, обеспеченное новыми сервисами БД, – это возможность «переключить внимание» LLM с глобальной среды на узкий терминологический ресурс с подбором более адекватной лексики.

В работе предложен подход к организации запросов на ЕЯ, использующий сервисы, реализованные в графовой БД AllegroGraph. Семантический ресурс, на котором изучались возможности и проблемы предложенного подхода, представлял собой онтологическую модель «Термаль», выполняющую роль управляющей надстройки и метаданных БД по теплофизическим свойствам вещества. Особенность ресурса – это богатый объем терминологии по классам веществ, их физическим свойствам и взаимосвязям при активном использовании двух языков, русского и английского.

После загрузки онтологической модели на облачный вариант БД AllegroGraph была проведена серия экспериментов по проверке релевантности ответов на переданные запросы. В целом проверка выполнения запросов к онтологии на русском и английском языках выявила работоспособность сервисов графовой БД AllegroGraph. Наряду с этим были зафиксированы явные различия в полноте ответов при запросах на различных языках. Анализ этих различий позволил выявить неполноту в представлении аннотационных свойств онтологии, а проведенная коррекция позволила улучшить совпадение ответов при разных языках запросов. В дальнейшем, помимо поиска в хранилище онтологий, планируется проведение организации аналогичного поиска в хранилище фактографических данных (текстов, таблиц, рисунков), размещенных на документно-ориентированной БД NoSQL типа Mongo DB.

Благодарности

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (Государственное задание № 075-00269-25-00).

СПИСОК ЛИТЕРАТУРЫ

1. *Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А.* Типология материаловедческих данных // Научно-техническая информация. Сер. 2. 2023. № 6. С. 25–39.

2. *Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Косинов А.В.* О представлении и оценке научных данных числового и нечислового типа при проведении исследований по свойствам материалов // Научно-техническая информация. Сер. 2. 2023. № 2. С. 8–16.

3. *Woods W.A.* Semantics and quantification in natural language question answering // Advances in computers. N.Y. etc.: Acad. Press, 1978. Vol. 17. P. 1–87. https://web.stanford.edu/class/linguist289/woods.pdf

4. *Бородин Д.С., Строганов Ю.В.* К задаче составления запросов к базам данных на естественном языке // Новые информационные технологии в автоматизированных системах: материалы 19-го научно-практического семинара. М.: ИПМ им. М.В. Келдыша, апрель 2016. С. 119–125.

5. *Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В.* Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. М.: МИЭМ, 2011. 272 с.

6. *Бородин Д.С., Строганов Ю.В., Волкова Л.Л., Рудаков И.В., Просуков Е.А.* Транслятор запросов на ограниченном естественном языке в запросы к реляционным базам данных // Системный администратор. 2019. Выпуск №01-02. С. 194–195.

7. *Посевкин Р.В.* Применение семантической модели базы данных при реализации естественно-языкового пользовательского интерфейса // Научно-технический вестник информационных технологий, механики и оптики. 2018. Том 18. №2. С. 262–267.

8. *Mikolov T., et al.* Distributed representations of words and phrases and their compositionality // Proc. 26th Int. Conf. on Neural Information Processing Systems. 2013. P. 3111–3119.

9. *Pennington J., et al*. Glove: Global vectors for word representation // Proc. Conf. Empirical Methods in Natural Language Processing. 2014. P. 1532–1543.

10. *Kenton J.D.M.-W. C., Toutanova L.K.* Bert: Pre-training of deep bidirectional transformers for language understanding // Proc. Conf. of North American Chapter of Association for Computational Linguistics. 2019. P. 4171–4186.

11. *Hafsa Shareef Dar, M. Ikramullah Lali, Khalid Mahmood Malik, Syed Ahmad Chan Bukhari.* Frameworks for Querying Databases Using Natural Language: A Literature Review. arXiv preprint. 2019. URL: https://arxiv.org/abs/1909.01822

12. *Baig Muhammad Shahzaib, et al.* Natural Language to SQL Queries: A Review Original Article // International Journal of Innovations in Science & Technology. 2022. Vol. 4. Issue 1. P. 147–162.

13. *Tao Yu, et al.* Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. arXiv preprint. 2018.
URL: https://arxiv.org/abs/1809.08887

14. *Manning C.D*. Human language understanding & reasoning // Daedalus 2022. Vol.151. Issue 2. P. 127–138.

15. *Meyer Jesse G., et al.* ChatGPT and large language models in academia: opportunities and challenges. // BioData Mining. 2023. Vol. 16. Art. numb. 20.

16. Microsoft Copilot в Azure с базой данных SQL Azure. URL: https://learn.microsoft.com/ru-ru/azure/azure-sql/copilot/copilot-azure-sql-overview?view=azuresql

17. MongoDB Query Generator using OpenAI.
URL: https://www.mongodb.com/docs/compass/current/query-with-natural-language/#std-label-compass-query-natural-language

18. Lower your Large Language Model costs with Graphwise GraphDB. URL: https://www.ontotext.com/blog/lower-your-llm-costs-with-graphwise-graphdb/

19. AllegroGraph 8.4.0 LLM Embed Specification.
URL: https://franz.com/ agraph/support/documentation/llmembed.html

20. Stardog Voicebox FAQ: How LLM, Generative AI, and Knowledge Graphs are the Future of Data Management.
URL: https://www.stardog.com/blog/stardog-voicebox-faq-how-llm-generative-ai-and-knowledge-graphs-are-the-future-of-data-management/

21. *Трахтенгерц М.С.* Технология подготовки информации для баз данных в обменном формате ISO 2709 // Научно-техническая информация. Сер. 2. 2006. № 7. С. 28–31.

QUERIES TO NON-RELATIONAL DATA USING NATURAL LANGUAGE BASED ON A LARGE LANGUAGE MODEL

A. O. Erkimbaev[1] [0000-0002-5239-2208], V. Yu. Zitserman[2] [0000-0003-3327-3139], G. A. Kobzev[3] [0000-0001-9987-1823]

[1-3]*Joint Institute for High Temperatures, RAS, 125412, Moscow, Russia*

[1]adilbek@jiht.ru, [2]vz1941@mail.ru, [3]gkbz@mail.ru

Abstract

The main purpose of this work is to explore new opportunities for organizing natural language queries in scientific local databases that are not relational. A brief review of recent research shows that there has been an active introduction of natural language queries into databases of various types, and the use of machine learning methods, such as neural algorithms, is noted. The widespread use of large language models in the last two years for query generation in various language settings and fields of expertise has been demonstrated. A study has been conducted to explore the potential of the AllegroGraph graph database in using large language models for natural language search. The functionality of the database has been examined using the example of a metadata system for thermophysical properties in the form of the "Thermal" domain ontology. Testing search queries in a bilingual (English and Russian) database environment has revealed some general problems that can be overcome, and it gives us good hope for the future application of new services using large language models.

Keywords*: natural language query, large language model, embedding, non-relational databases, graph database, domain ontology.*

REFERENCES

1. *Erkimbaev A.O., Zitserman V.Iu., Kobzev G.A.* Tipologiia materialovedcheskikh dannykh // Nauchno-tekhnicheskaia informatsiia. Ser. 2. 2023. № 6. S. 25–39.

2. *Erkimbaev A.O., Zitserman V.Iu., Kobzev G.A., Kosinov A.V.* O predstavlenii i otsenke nauchnykh dannykh chislovogo i nechislovogo tipa pri provedenii issledovanii

po svoistvam materialov // Nauchno-tekhnicheskaia informatsiia. Ser. 2. 2023. № 2. S. 8–16.

3. *Woods W.A.* Semantics and quantification in natural language question answering. // Advances in computers. N.Y. etc.: Acad. Press, 1978. Vol. 1 7. P. 1–87. https://web.stanford.edu/class/linguist289/woods.pdf

4. *Borodin D.S., Stroganov Iu.V.* K zadache sostavleniia zaprosov k bazam dannykh na estestvennom iazyke // Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh: materialy 19 nauchno-prakticheskogo seminara. M.: IPM im. M.V. Keldysha, aprel 2016. P. 119–125.

5. *Bolshakova E.I., Klyshinskii E. S., Lande D.V., Noskov A.A., Peskova O.V., Iagunova E.V.* Avtomaticheskaia obrabotka tekstov na estestvennom iazyke i kompiuternaia lingvistika: uchebnoe posobie. M.: MIEM, 2011. 272 s.

6. *Borodin D.S., Stroganov Iu.V., Volkova L.L., Rudakov I.V., Prosukov E.A.* Transliator zaprosov na ogranichennom estestvennom iazyke v zaprosy k reliatsionnym bazam dannykh // Sistemnyi administrator. 2019. Vypusk №01-02. S. 194–195.

7. *Posevkin R.V.* Primenenie semanticheskoi modeli bazy dannykh pri realizatsii estestvenno-iazykovogo polzovatelskogo interfeisa // Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki. 2018. Tom 18. № 2. S. 262–267.

8. *Mikolov T., et al.* Distributed representations of words and phrases and their compositionality // Proc. 26th Int. Conf. on Neural Information Processing Systems. 2013. P. 3111–3119.

9. *Pennington J., et al.* Glove: Global vectors for word representation // Proc. Conf. Empirical Methods in Natural Language Processing. 2014. P. 1532–1543.

10. *Kenton J.D.M.-W. C., Toutanova L.K.* Bert: Pre-training of deep bidirectional transformers for language understanding // Proc. Conf. of North American Chapter of Association for Computational Linguistics. 2019. P. 4171–4186.

11. *Hafsa Shareef Dar, M. Ikramullah Lali, Khalid Mahmood Malik, Syed Ahmad Chan Bukhari*. Frameworks for Querying Databases Using Natural Language: A Literature Review. 2019. P. 1–18. arXiv preprint.
URL: https://arxiv.org/abs/1909.01822

12. *Baig Muhammad Shahzaib, et al.* Natural Language to SQL Queries: A Review Original Article // International Journal of Innovations in Science &

Technology. 2022. Vol. 4. Issue 1. P. 147–162.

13. *Tao Yu, et al.* Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. arXiv preprint. 2018. URL: https://arxiv.org/abs/1809.08887

14. *Manning C.D.* Human language understanding & reasoning // Daedalus 2022. Vol. 151. Issue 2. P. 127–138.

15. *Meyer Jesse G., et al.* ChatGPT and large language models in academia: opportunities and challenges // BioData Mining 2023. Vol. 16. Art. numb. 20.

16. Microsoft Copilot в Azure с базой данных SQL Azure. URL: https://learn.microsoft.com/ru-ru/azure/azure-sql/copilot/copilot-azure-sql-overview?view=azuresql

17. MongoDB Query Generator using OpenAI. URL: https://www.mongodb.com/docs/compass/current/query-with-natural-language/#std-label-compass-query-natural-language

18. Lower your Large Language Model costs with Graphwise GraphDB. URL: https://www.ontotext.com/blog/lower-your-llm-costs-with-graphwise-graphdb/

19. AllegroGraph 8.4.0 LLM Embed Specification. URL: https://franz.com/agraph/support/documentation/llmembed.html

20. Stardog Voicebox FAQ: How LLM, Generative AI, and Knowledge Graphs are the Future of Data Management. URL: https://www.stardog.com/blog/stardog-voicebox-faq-how-llm-generative-ai-and-knowledge-graphs-are-the-future-of-data-management/

21. *Trakhtengerts M.S.* Tekhnologiia podgotovki informatsii dlia baz dannykh v obmennom formate ISO 2709 // Nauchno-tekhnicheskaia informatsiia. Ser. 2. 2006. № 7. S. 28–31.

СВЕДЕНИЯ ОБ АВТОРАХ

ЕРКИМБАЕВ Адильбек Омирбекович – старший научный сотрудник лаборатории теплофизических баз данных Объединенного института высоких температур (ОИВТ РАН), кандидат техн. наук. Область научных интересов: теплофизика, теплофизические свойства веществ, технологии баз данных.

Adilbek Omirbekovich ERKIMBAEV – Senior Researcher at the Thermophysical Databases Laboratory, Joint Institute for High Temperatures (JIHT RAS), PhD. Research interests: thermophysics, thermophysical properties of substances, database technologies.

email: adilbek@jiht.ru

ORCID: 0000-0002-5239-2208

ЗИЦЕРМАН Владимир Юрьевич – ведущий научный сотрудник лаборатории теплофизических баз данных Объединенного института высоких температур (ОИВТ РАН), кандидат физ.-мат. наук. Область научных интересов: теплофизика, химическая физика, технологии баз данных.

Vladimir Yurievich ZITSERMAN – Leading Researcher at the Thermophysical Databases Laboratory, Joint Institute for High Temperatures (JIHT RAS), Ph.D. in Physico-mathematical Sciences. Research interests: thermophysics, chemical physics, database technologies.

email: vz1941@mail.ru

ORCID: 0000-0003-3327-3139

КОБЗЕВ Георгий Анатольевич – главный научный сотрудник лаборатории теплофизических баз данных Объединенного института высоких температур (ОИВТ РАН), доктор физ.-мат. наук. Область научных интересов: теплофизика, физика неидеальной плазмы, систематизация научных данных

George Anatolyevich KOBZEV – Principal Research Scientist at the Thermophysical Databases Laboratory, Joint Institute for High Temperatures (JIHT RAS), DSc (Phys), Research interests: thermophysics, the physics of non-ideal plasmas, scientifical data categorization.

email: gkbz@mail.ru

ORCID: 0000-0001-9987-1823

Материал поступил в редакцию 12 декабря 2025 года