ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 004.682 *Научная статья* DOI: 10.18287/2223-9537-2025-15-2-239-248

Использование онтологий для контекстуализации запросов к большим языковым моделям

© 2025, П.А. Ломов

Кольский научный центр Российской академии наук,
Институт информатики и математического моделирования им. В.А. Путилова, Апатиты, Россия

Аннотация

Применение больших языковых моделей стало распространённым явлением в вопросно-ответных и диалоговых системах. Для этого модель должна быть предварительно обучена на подготовленных текстовых данных, что позволяет ей с высокой вероятностью генерировать корректные реплики в диалоге с пользователем. Однако качество ответов снижается, если вопросы начинают касаться предметов, процессов и явлений, которые в меньшей степени описаны в текстах, использованных для обучения модели. Для этого данные, являющиеся новыми для модели, передаются ей вместе с пользовательским запросом в виде контекста, который обычно формируется с использование векторной базы данных текстовых фрагментов. В статье предлагается использование в качестве источника контекстных данных вместо векторной базы данных онтологии предметной области. Онтологии снабжаются лексическим представлением формализованной в них терминологической системы для идентификации релевантного пользовательскому запросу онтологического фрагмента и трансформации его в естественно-языковой текст формируемого контекста. Это позволяет уменьшить объём текста ответа и повысить степень его семантического соответствия пользовательскому запросу. В статье рассматриваются минимальные требования к структуре лексического представления онтологии: наличие естественно-языковых наименований, их форм для понятий и отношений, а также их лексических значений. Применение предложенного подхода показано на примере получения ответа на вопрос по научным статьям с использованием большой языковой модели. Обсуждаются преимущества и недостатки предложенного подхода.

Ключевые слова: *онтология, большая языковая модель, запрос, контекст, генерация ответа.*

Цитирование: *Ломов П.А.* Использование онтологий для контекстуализации запросов к большим языковым моделям. *Онтология проектирования*. 2025. Том 15, №2(56). С.239-248. DOI: 10.18287/2223-9537-2025-15-2-239-248.

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Введение

Одним из перспективных направлений искусственного интеллекта является создание и применение больших языковых моделей (БЯМ) [1, 2]. С их помощью удаётся решать задачи, связанные с обработкой естественно-языковых (ЕЯ) текстов, включая их генерацию, классификацию, определение тем, автоматический перевод и др. Существенные результаты получены при использовании БЯМ для создания вопросно-ответных и диалоговых систем [3,4]. Для этого модель должна быть предварительно обучена на огромном объёме подготовленных текстовых данных, что позволяет ей генерировать корректные реплики в диалоге с пользователем [5]. Однако в таком сценарии использования качество ответов модели, как правило, падает, если вопросы пользователя начинают касаться предметов, процессов и явлений, которые в сравнительно меньшей степени описаны в текстах, использованных для её обуче-

ния. Для решения этой проблемы можно провести дообучение модели на текстах, отражающих необходимую специфику предметных областей (ПрО) и/или прикладных задач, но это может потребовать существенных аппаратных и временных ресурсов.

Широкое распространение получил подход к обеспечению учёта моделью новой информации на основе генерации ответа с использованием результатов поиска (*Retrieval-Augmented Generation, RAG*) [6]. Его суть заключается в автоматическом получении дополнительных текстовых данных, с которыми предобученая модель не знакома, и их добавлении в запрос пользователя. Упомянутые данные должны быть релевантны запросу, включающему указание модели на необходимость их учёта при генерации ответа. Таким образом, *RAG* предполагает использование расширенного запроса к БЯМ (промпта, *prompt*), включающего, помимо текста формулировки пользовательского вопроса или задания, некоторый контекст – данные, на основе которых модель должна сгенерировать текст ответа.

RAG можно рассматривать как разновидность контекстного обучения [7], где контекст формируется автоматически путём извлечения текстовых фрагментов из базы данных (БД). В качестве хранилища для *RAG* часто используются векторные БД, которые позволяют получать подходящие для добавления в контекст запроса тексты по близости их векторных представлений к векторному представлению текста запроса. При использовании векторной БД необходимо представить в ограниченном по размеру контексте информацию о всех понятиях из пользовательского запроса. В противном случае ответ модели может быть неполным или абстрактным. Для уменьшения размера контекста можно разделить тексты на фрагменты при добавлении в БД и/или использовать для поиска не весь текст запроса, а только именные группы, которые обозначают значимые понятия и отношения между ними. Таким способом можно дополнительно структурировать информацию, которую модель будет использовать. В общем случае чёткие правила структурирования сложно сформулировать без семантического анализа текстов и обнаружения в них понятий и отношений ПрО.

В данной работе предлагается использование существующей онтологии ПрО вместо векторной БД для контекстуализации запросов к БЯМ в рамках *RAG*.

1 Обзор основных подходов

RAG представляет собой перспективный подход к преодолению присущих БЯМ ограничений, связанных с их приоритетной ориентацией на знания, представленные в виде ЕЯ текстов, и с затруднительной актуализацией этих знаний. Это и определяет большое количество исследований [8], посвящённых реализациям *RAG*, направленным на ускорение получения контекстных данных, повышение их релевантности пользовательскому запросу, обеспечение использования знаний, представленных в разных модальностях, и упрощение процесса развёртывания и использования *RAG*.

В работе [6] впервые был представлен подход к использованию БЯМ, включающий комбинирование процессов извлечения информации из внешних баз знаний с генерацией текста. *RAG* включает две компоненты: поисковый модуль, отвечающий за извлечение релевантных фрагментов текста из БД; генератор, формирующий ответ с учётом извлечённой информации. Данная работа определила основную архитектуру систем, реализующих *RAG*, и послужила отправной точкой для последующих исследований в этой области.

В работе [9] поисковый модуль был использован не на этапе обработки пользовательского запроса, а на этапе предварительного обучения модели кодировщика *BERT* (*Bidirectional Encoder Representations from Transformers*) для подбора документов, представляющих знания о ПрО. Модель кодировщика обычно предобучается на корпусе общеязыковых текстов, из которого ей последовательно передаются автоматически размеченные предложения. Соглас-

но [9] поисковый модуль для каждого «обычного» предложения пытается найти семантически близкое предложение из корпуса текстов по ПрО. В результате кодировщик получает обучающий пример, состоящий из двух предложений. Это позволяет осуществить обучение на комбинации примеров из общеязыкового и специализированного наборов и обеспечить возможность генерации текстов с учётом знаний ПрО.

В работе [10] описано применение *RAG* в задаче ведения голосового диалога пользователя с системой. В этом случае используются ранее сохранённые фрагменты диалогов, с которыми сопрягаются характеризующие их семантические и стилевые векторы, для генерации системой ответных речевых реплик с учётом общего смысла диалога и его языкового стиля.

Близкими к подходу, который рассматривается в данной работе, можно отнести подходы с использованием графов (*Graph RAG*) [11], в которых языковая модель при генерации текста может обращаться к структурированным знаниям, представленным в виде графа. Это позволяет ей использовать информацию о взаимосвязях между различными сущностями ПрО, повышая релевантность и точность генерируемых ею ответов.

Для представления данных ПрО в виде графа, как правило, используются модели: *LPG* (*Label Property Graph*) или *RDF* (*Resource Description Framework*). Для этого требуется приведение знаний ПрО в графовый вид в графовой БД, а также определение в ней индексов различных типов для ускорения выполнения запросов. Наличие графового представления знаний ПрО обеспечивает гибкость поискового модуля, в котором могут использоваться простые эвристики и традиционные алгоритмы поиска по графу [12, 13], БЯМ для трансляции ЕЯ вопросов в запросы к графовой БД на специализированном языке [14], а также графовая нейронная сеть для кодирования структуры графа и выбора подходящих элементов (узлы, пути, подграфы) на основании их сходства с пользовательским запросом [15].

2 Применение онтологии в *RAG*-технологии

Онтология ПрО представляет её понятийную систему, описанную на некотором формальном языке, что делает возможным её использование для обработки данных в информационных системах. При разработке онтологии важно обеспечить интерсубъективность [16] явленных в ней знаний о ПрО. Именно соблюдение этого требования делает онтологию разделяемой формальной спецификацией концептуализации [17], т.е. концептуализации, являющейся результатом консенсуса некоторого профессионального сообщества.

Данное свойство онтологии позволяет рассматривать её как истинную в некотором отношении основу для структурирования предметных ЕЯ текстов и соотнесения их фрагментов с онтологическими понятиями. Представленные в онтологии ПрО межпонятийные отношения и атрибуты понятий могут быть транслированы в ЕЯ предложения, смысл которых может быть близок к смыслу предложений из текстов ПрО, содержащих эти понятия.

Использование онтологии в *RAG* состоит в формировании ЕЯ контекста пользовательского запроса из фрагментов текстов, сопряжённых с онтологическими понятиями, релевантными этому запросу, а также из ЕЯ представлений (лексикализаций) их связей с другими элементами онтологии (понятиями, атрибутами, значениями). Полнота и корректность формируемого контекста в этом случае зависят от полноты и детализации лексического представления понятийной системы онтологии [18]. В данном случае такое представление должно включать для каждого понятия его наименование и текстовый фрагмент с лексическим значением. Для отношений и атрибутов необходимо представить наименования, дополнив их возможными формами для обеспечения возможности построения грамматически корректных ЕЯ предложений, представляющих различные варианты их употребления.

Для иллюстрации можно рассмотреть лексическое представление в онтологии понятия «Составной онтологический паттерн (ОП) содержания» (*CompositeOntologyDesignPattern*) (см. рисунок 1) и отношения «содержит» (*has_part*) (см. рисунок 2), которое используется в связанной с ним *OWL* аксиоме:

CompositeOntologyDesignPattern *EquivalentTo:* ***ContentOntologyDesignPattern*** *and (has_part some* ***ContentOntologyDesignPattern****).*

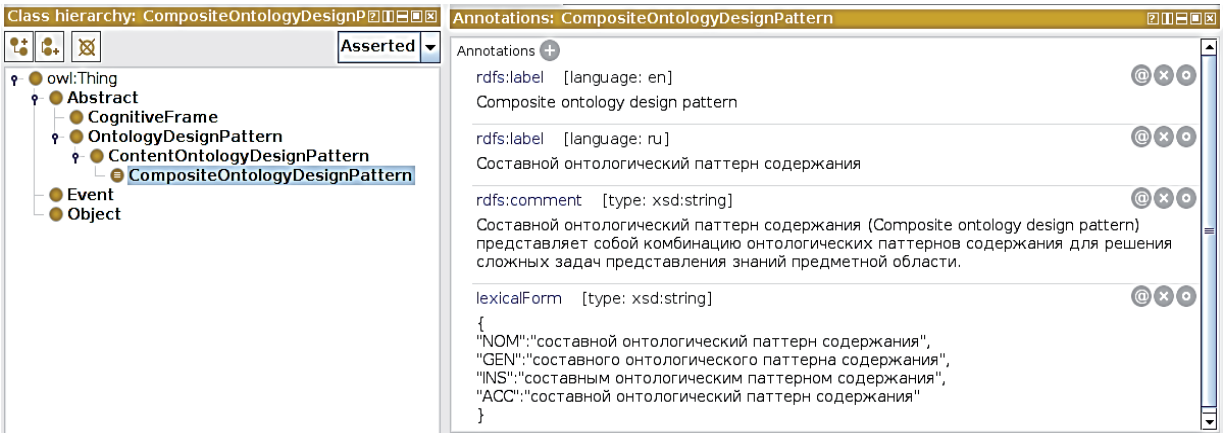

Рисунок 1 – Фрагмент онтологии с лексическим представлением понятия онтологии «Составной онтологический паттерн содержания» (*Composite ontology design pattern*)

Понятие имеет несколько аннотаций, включающих ЕЯ наименования (*rdfs*:*label*) и определения (*rdfs*:*comment*) из [19], и несколько форм употребления (*lexicalForm*) в различных падежах. Отношение имеет несколько аннотаций: наименование и возможные формы наименований понятий (*domainLexicalForm*, *rangeLexicalForm*), которые оно может связывать.

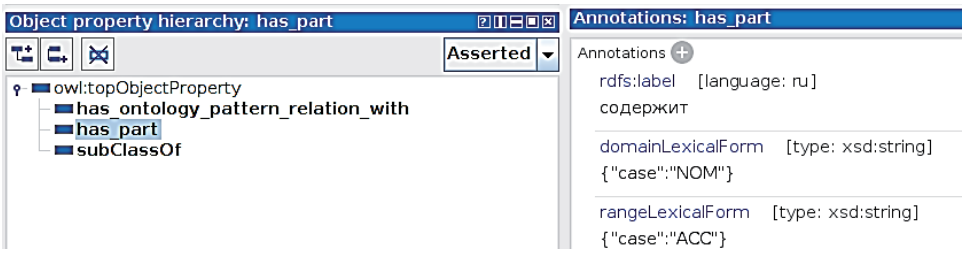

Рисунок 2 - Фрагмент онтологии с лексическим представлением отношения «содержит» (*has_part*)

3 Процедура контекстуализации запросов к большим языковым моделям

Процедура формирования контекста с использованием онтологии ПрО для пользовательского запроса к БЯМ состоит из двух этапов: идентификация фрагмента онтологии, релевантного запросу пользователя; трансляция онтологического фрагмента в ЕЯ текст, составляющий содержание контекста.

Можно рассмотреть применение данной процедуры на примере формирования онтологического контекста для пользовательского запроса и последующую передачу их БЯМ в виде промпта. В качестве источника данных для получения контекста использовалась онтология, фрагменты которой представлены на рисунках 1 и 2, содержащая понятия из статей [19, 20]. В качестве БЯМ использовался *GigaChat-Pro* (https://giga.chat/) от компании Сбер (по состоянию на 20.08.2024).

Рассматривался пользовательский запрос: «Как отношения между ОП используются для формирования когнитивных фреймов (КФ)?» Ответ на него содержится в [19, 20].

В [20] рассматривалась концепция КФ, которая представляет собой формализованное описание визуализации некоторой точки зрения на понятие. В качестве «точек зрения» применительно к онтологиям рассматривались ОП содержания, которые представляют собой мини-онтологии для описания типовых положений вещей в ПрО (например, часть-целое, участие в событии, исполнитель роли). Таким образом, ОП является основой для построения КФ. В [19] рассматривалось объединение нескольких ОП для получения составного ОП, который может быть использован для описания сложной ситуации в ПрО (например, участие сущностей в событии в определённых ролях). Однако такое комбинирование ОП возможно в том случае, если между ними существуют отношения (специализирует, имеет часть, соотносится).

Правильный ответ на приведённый запрос должен включать указание на то, что именно отношения позволяют получить составной ОП, на основе которого можно построить КФ. Сформулировать его можно следующим образом: «Отношения позволяют получить составной ОП, на основе которого можно построить КФ».

В качестве формальной оценки соответствия ответов, сгенерированных БЯМ, и правильного используется косинусная близость между их контекстуализированными векторными представлениями (*contextualized word embeddings*) [21]. Для получения таких представлений могут быть использованы кодировщик *BERT*, предобученный на текстах требуемого языка, или модель *ELMo* (*Embeddings from Language Models*) [22]. В рассматриваемом случае использовался кодировщик *ruBert-base* [23], который был дообучен на [19, 20] для учёта специфики лексики и грамматических конструкций, используемых в онтологии. Для получения шкалированых значений косинусной близости (ШКБ) генерируемых БЯМ ответов (приведения значений ШКБ к единичному интервалу) необходимо определить минимально и максимально возможную близость. Для этого задаются несколько вариантов правильного ответа.

1) отношения помогают сформировать сложный ОП, который служит основой для построения КФ.
2) благодаря отношениям создаётся составной ОП, служащий базой для КФ.
3) посредством отношений формируется комбинированный ОП, используемый для конструирования КФ.
4) отношения позволяют получить составной ОП. Этот паттерн используется для построения КФ.

Найденное среднее значение косинусной близости между векторными представлениями данных вариантов и векторными представлением правильного ответа составляет 0,942. Это значение принято за близость, соответствующую наиболее точным и полным ответам. Аналогичным образом определена близость для максимально плохих ответов – 0,991. В качестве таковых произвольным образом выбрано несколько предложений из [19, 20], включающих некоторые понятия из рассматриваемого вопроса, однако не отвечающих на него.

Формирование контекста начинается с определения понятий онтологии, наиболее близких к запросу пользователя. В качестве метрики применена косинусная близость векторных представлений вопроса пользователя и лексического значения понятия. Для получения таких представлений использовался кодировщик *BERT*. Были получены оценки семантической близости между запросом пользователя и всеми понятиями онтологии. В качестве выходных были выбраны три понятия, имеющие наибольшую близость: «ОП содержания» (*ContentOntologyDesignPattern*), «Когнитивный фрейм» (*CognitiveFrame*) и «Составной ОП содержания» (*CompositeOntologyDesignPattern*).

Второй этап формирования контекста запроса включает извлечение фрагмента онтологии, относящегося к выбранным понятиям. Для этого из онтологии извлекаются логические выражения (*OWL* аксиомы), которые имеют в левых частях данные понятия. Например, для понятия *CompositeOntologyDesignPattern* такими *OWL* аксиомами являются:

CompositeOntologyDesignPattern *SubClassOf* ***ContentOntologyDesignPattern***
CompositeOntologyDesignPattern *EquivalentTo* ***ContentOntologyDesignPattern*** *and (has_part some* ***ContentOntologyDesignPattern****)*

Далее производится конвертация найденных *OWL* аксиом в ЕЯ предложения с помощью лексического представления онтологии, которое содержит ЕЯ наименования понятий и отношений, а также их формы. Данный процесс рассматривается на примере *OWL* аксиомы:

CompositeOntologyDesignPattern *equivalentTo* ***ContentOntologyDesignPattern*** *and (has_part some* ***ContentOntologyDesignPattern****)*

Понятия *CompositeOntologyDesignPattern, ContentOntologyDesignPattern* используются в данной аксиоме, как носитель и значение свойства *has_part*, лексическое представление которого предписывает использовать для них соответственно формы именительного {*case: NOM*} и винительного {*case: ACC*} падежей (рисунок 1), т.е. «Составной ОП содержания» и «ОП содержания». Свойство *has_part* имеет наименование «содержит». Таким образом, ЕЯ предложение для рассматриваемой аксиомы имеет вид:

Составной ОП содержит ОП содержания

Подобным образом с использованием лексического представления для свойства *subClassOf* (рисунок 3) производится конвертация *SubClassOf* аксиом, используемых для определения иерархии наследования.

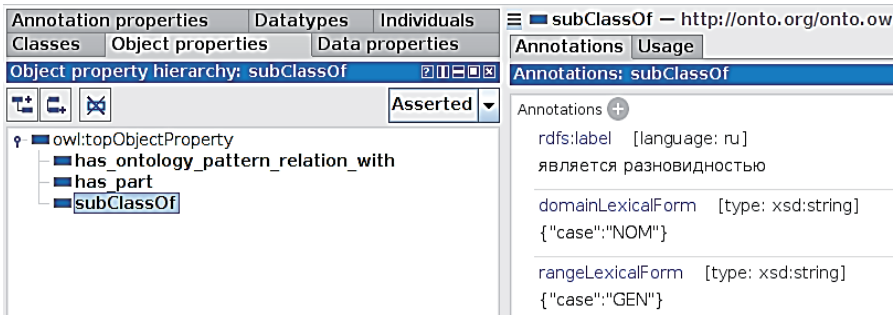

Рисунок 3 - Фрагмент онтологии с лексическим представлением отношения наследования (*subClassOf*)

Например, для *OWL* аксиомы:

CompositeOntologyDesignPattern *SubClassOf* ***ContentOntologyDesignPattern***

соответствующее ЕЯ предложение имеет вид:

Составной ОП содержания является разновидностью ОП содержания

Полученные таким образом ЕЯ предложения добавляются в контекст запроса наряду с лексическими значениями понятий, представленный в виде *rdfs:comment* аннотаций в онтологии. Итоговый промпт для БЯМ, включающий вопрос и его контекст, имеет вид:

Ответь на вопрос только с помощью указанного контекста: *Вопрос:* Как отношения между ОП используются для формирования КФ? *Контекст:* «ОП содержания (*Content ontology design patterns, CDP*) описывает вариант представления знаний ПрО в виде фрагментов онтологий, т.е. нескольких классов онтологии, связанных отношениями. Для каждого паттерна содержания задаётся обобщённое определение ситуации (*General Use Case, GUC*), в которой необходимо его применять. Например, участие в событии, исполнение роли, наличие частей у объекта и другие… ОП содержания является разновидностью ОП проектирования. Составной ОП содержания является разновидностью ОП содержания. Составной ОП содержания содержит ОП содержания».

Были получены следующие ответы, имеющие примерно одинаковое значение ШКБ (т.е. формально они одинаковые с точки зрения близости к правильному ответу):

Ответ 1. ШКБ= 0.100. Отношения между ОП используются для формирования КФ следующим образом:

На основе одного или нескольких ОП содержания, которые представляют варианты представления знаний ПрО в виде фрагментов онтологий, формируются КФ. Эти КФ визуализируют фрагменты онтологии и передают экспертам знания о понятиях, включая их содержание и визуальные образы. КФ могут быть сформированы как из простых ОП содержания, так и из составных, которые представляют комбинации паттернов для решения более сложных задач представления знаний.

Ответ 2. ШКБ= 0.103. Отношения между ОП используются для формирования КФ следующим образом:

1. ОП содержания (*CDP*) представляют варианты представления знаний в виде фрагментов онтологий. Эти фрагменты состоят из классов онтологии, связанных отношениями.

2. Для каждого паттерна содержания задаётся обобщённое определение ситуации (*GUC*), которая определяет контекст использования данного паттерна.

3. На основе этих определений ситуаций задаются квалификационные вопросы, которые могут быть преобразованы в запросы к онтологии. Ответы на эти вопросы гарантированно будут получены при реализации соответствующих паттернов…

6. Составные ОП содержания представляют собой комбинации ОП для решения сложных задач представления знаний. Они содержат один или несколько паттернов содержания.

Полученные варианты ответов абстрактны и не содержат ожидаемую информацию, обозначенную в приведённом правильном ответе. Для выяснения причины можно проверить, достаточно ли в онтологии информации о понятиях, используемых в пользовательском запросе. Сделать это проще, чем в случае использования векторной БД, ввиду того, что в онтологии информация о ПрО распределена по её элементам (классам, экземплярам, свойствам и др.), которые образуют интуитивно понятную эксперту систему.

Для понятия *CompositeOntologyDesignPattern* (Составной ОП содержания) нет упоминания о его связи с отношениями ни в его ЕЯ определении, ни в *OWL* аксиомах (рисунок 1). Для включения отсутствующей информации в контекст запроса необходимо добавить её в онтологию в виде аксиом, формально определяющих понятие «Отношение между ОП содержания», или дополнить ею ЕЯ определение.

Воспользовавшись вторым способом (ввиду его относительной простоты), можно вставить соответствующий текстовый фрагмент (рисунок 4).

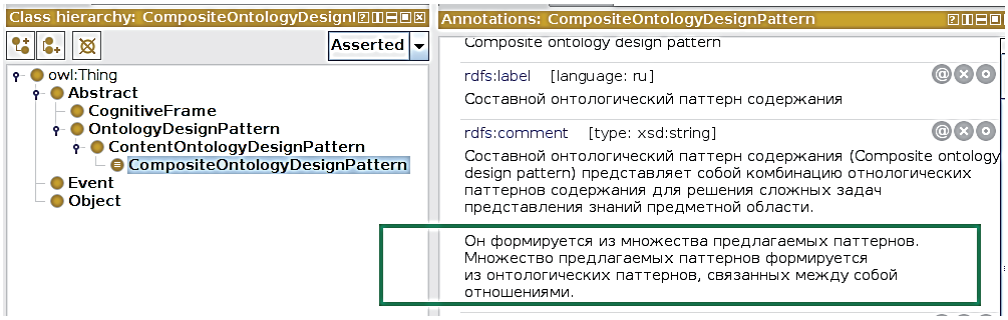

Рисунок 4 - Расширенное определение понятия «Составной онтологический паттерн содержания» (CompositeOntologyDesignPattern)

Упоминание отношений попадает в контекст запроса, и БЯМ даёт следующий ответ:

Отношения между ОП используются для формирования составных ОП содержания, которые в свою очередь служат основой для создания КФ. КФ представляют собой визуализированные фрагменты онтологии, передающие знания о понятиях. Они включают в себя содержание, сформированное на основе отношений «понятие-отношение-понятие», и визуальный образ. Таким образом, отношения между ОП играют ключевую роль в структурировании и представлении знаний в форме КФ.

ШКБ данного ответа составляет 0.24, что превышает оценки полученных ранее ответов в два раза и свидетельствует о том, что данный ответ уже в большей степени соответствует правильному. Если его рассмотреть содержательно, то можно обнаружить, что его первое предложение включает ожидаемый ответ: «Отношения позволяют получить составной ОП, на основе которого можно построить КФ».

Заключение

В работе предложен подход к применению технологии *RAG*, ориентированной на построение контекста для пользовательского запроса к БЯМ на основе онтологии ПрО вместо векторной БД. Рассмотрено лексическое представление понятийной системы онтологии, поз-

воляющее интерпретировать её фрагменты в виде ЕЯ текстов. Это обеспечивает возможность формирования из них контекста пользовательского запроса с целью предоставления БЯМ данных для генерации ответа.

За счёт использования онтологии обеспечивается больший контроль над получаемым контекстом, чем в случае использования векторной БД, т.к. его содержимое определяется элементами понятийной системы онтологии и их лексическим представлением, которые могут быть изменены экспертом ПрО.

Предложенный подход может быть использован для отладки онтологии экспертом ПрО, незнакомым со специализированными языками запросов и приёмами онтологического моделирования, т.к. позволяет опрашивать онтологию с помощью ЕЯ вопросов с целью выявления в ней пробелов и неточностей.

Список источников

[1] ***Zhao Z., Zhouю K., Li J., Tang T.*** A Survey of Large Language Models. 2024. DOI: 10.48550/arXiv.2303.18223.

[2] ***Minaee S., Mikolov T., Nikzad N.*** Large Language Models: A Survey. 2024. DOI: 10.48550/arXiv.2402.06196.

[3] ***Li Z., Peng J., Wang Y.*** ChatSOP: An SOP-Guided MCTS Planning Framework for Controllable LLM Dialogue Agents. 2025. DOI: 10.48550/arXiv.2407.03884.

[4] ***Roller S.E., Dinan E., Goyal N.*** Recipes for building an open-domain chatbot. 2020. DOI: 10.48550/arXiv.2004.13637.

[5] ***Zhang H., Li W.W., Chen R.L.*** LLM-Enhanced Dialogue Management for Full-Duplex Spoken Dialogue Systems. 2025. DOI: 10.48550/arXiv.2502.14145.

[6] ***Lewis P., Perez E., Piktus A.*** Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2021. DOI: 10.48550/arXiv.2502.14145.

[7] ***Sia S., Duh K.*** In-context Learning as Maintaining Coherency: A Study of On-the-fly Machine Translation Using Large Language Models. 2023. DOI: 10.48550/arXiv.2305.03573.

[8] ***Arslan M., Ghanem H., Munawar S.*** A Survey on RAG with LLMs. *Procedia Computer Science*. 2024. Vol.246. P.3781–3790. DOI: 10.1016/j.procs.2024.09.178.

[9] ***Guu K., Lee K.*** REALM: Retrieval-augmented language model pre-training. 2020. DOI: 10.48550/arXiv.2002.08909.

[10] ***Liu R., Jia Z., Bao F.*** Retrieval-Augmented Dialogue Knowledge Aggregation for expressive conversational speech synthesis. *Information Fusion*. 2025. Vol.118. P.102948. DOI: 10.48550/arXiv.2501.06467.

[11] ***Peng B., Zhu Y., Liu Y.*** Graph retrieval-augmented generation: a survey. 2024. DOI: 10.48550/arXiv.2408.08921.

[12] ***Yasunaga M., Ren H., Bosselut A.*** QA-GNN: Reasoning with language models and knowledge graphs for question answering. 2022. DOI: 10.48550/arXiv.2104.06378.

[13] ***Taunk D., Khanna L., Kandru P.*** GrapeQA: GRraph augmentation and pruning to enhance question-answering. 2023. DOI: 10.48550/arXiv.2303.12320.

[14] ***Zhang J., Zhang X., Yu J.*** Subgraph retrieval enhanced model for multi-hop knowledge base question answering // Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers). *Association for Computational Linguistics*, 2022. DOI: 10.18653/v1/2022.acl-long.396.

[15] ***Mavromatis C., Karypis G.*** GNN-RAG: Graph neural retrieval for large language model reasoning. 2024. DOI: 10.48550/arXiv.2405.20139.

[16] ***Боровик С.Ю.*** Онтологии, интерсубъективное управление и эвергетика В.А. Виттиха. *Онтология проектирования*. 2020. Том 10, №3. С.255–272. DOI: 10.18287/2223-9537-2020-10-3-255-272.

[17] ***Guarino N., Oberle D., Staab S.*** What is an ontology? // Handbook on ontologies. Springer, 2009. P.1–17. DOI: 10.1007/978-3-540-92673-3_0.

[18] ***Ломов П.А.*** Формирование лексического модуля прикладной онтологии для её обучения. *Онтология проектирования*. 2024. Том 13, №4. С.520–530. DOI: 10.18287/2223-9537-2023-13-4-520-530.

[19] ***Ломов П.А.*** Автоматизация синтеза составных онтологических паттернов содержания. *Онтология проектирования*. 2016. Том 6, №2. С.162–172. DOI: 10.18287/2223-9537-2016-6-2-162-172.

[20] ***Ломов П.А., Шишаев М.Г.*** Формирование когнитивных фреймов на основе онтологических паттернов для визуализации онтологий. *Информационные системы и технологии*. 2015. №6. С.12–22.

[21] ***Reimers N., Gurevych I.*** Sentence-BERT: Sentence embeddings using siamese BERT-networks. 2019. DOI: 10.48550/arXiv.1908.10084.
[22] ***Peters M.E., Neumann M., Iyyer M.*** Deep contextualized word representations. 2018. DOI: 10.48550/arXiv.1802.05365.
[23] ***Zmitrovich D., Abramov A., Kalmykov A.*** A family of pretrained transformer language models for russian. 2023. DOI: 10.48550/arXiv.2309.10931.

Сведения об авторе

Ломов Павел Андреевич, 1984 г. рождения, к.т.н., старший научный сотрудник Института информатики и математического моделирования имени В.А. Путилова Кольского научного центра РАН. Области научных интересов: представление знаний, онтологическое моделирование, семантическая сеть. AuthorID (РИНЦ): 8479-8320. Author ID (Scopus): 55350587100; ORCID: 0000-0002-0924-0188; Researcher ID (WoS): P-6627-2015. *palandlom@yandex.ru*.

Поступила в редакцию 02.02.2025, после рецензирования 10.03.2025. Принята к публикации 20.03.2025.

Scientific article DOI: 10.18287/2223-9537-2025-15-2-239-248

Using ontologies to contextualize queries to large language models

© 2025, P.A. Lomov

Kola Science Centre of the Russian Academy of Science,
Institute for Informatics and Mathematical Modeling named after V.A. Putilov, Apatity, Russia

Abstract

The use of large language models has become a common phenomenon in question-answering and dialog systems. For this, the model must be pre-trained on prepared text data, enabling it to generate highly probable correct responses in a dialog with the user. However, answer quality decreases when the questions pertain to objects, processes, or phenomena that are less described in the texts used to train the model. For this purpose, data that is new to the model is transferred to it along with the user query in the form of context, which is usually generated using a vector database of text fragments. The article proposes to use an ontology of a subject area as a source of contextual data instead of a vector database. Ontologies are supplied with a lexical representation of their formalized terminology system to identify an ontological fragment relevant to the user query and convert it into a natural language text of the formed context. This allows to reduce the response text volume while improving its semantic alignment with the user query. The article discusses the minimum structural requirements for the lexical representation of an ontology, including natural language names, their forms for concepts and relations, as well as their lexical meanings. The application of the proposed approach is shown through an example of obtaining an answer to a question on scientific articles using a large language model. The advantages and disadvantages of the proposed approach are discussed.

Keywords: *ontology, large language model, query, context, response generation*.

For citation: *Lomov PA*. Using ontologies to contextualize queries to large language models [In Russian]. *Ontology of designing*. 2025; 15(2): 239-248. DOI:10.18287/2223-9537-2025-15-2-239-248.

Conflict of interest: The author declares no conflict of interest.

List of figures

Figure 1 - Lexical representation of the ontology concept “Composite ontology design pattern”
Figure 2 - Lexical representation of the “has part” relation
Figure 3 - Lexical representation of the inheritance relation (subClassOf)
Figure 4 - Extended definition of the ontology concept “Composite ontology design pattern”

References

[1] ***Zhao Z, Zhouyu K, Li J, Tang T.*** A Survey of Large Language Models. 2024. DOI: 10.48550/arXiv.2303.18223.

[2] ***Minaee S, Mikolov T, Nikzad N.*** Large Language Models: A Survey. 2024. DOI: 10.48550/arXiv.2402.06196.

[3] ***Li Z, Peng J, Wang Y.*** ChatSOP: An SOP-Guided MCTS Planning Framework for Controllable LLM Dialogue Agents. 2025. DOI: 10.48550/arXiv.2407.03884.

[4] ***Roller SE, Dinan E, Goyal N.*** Recipes for building an open-domain chatbot. 2020. DOI: 10.48550/arXiv.2004.13637.

[5] ***Zhang H, Li WW, Chen RL.*** LLM-Enhanced Dialogue Management for Full-Duplex Spoken Dialogue Systems. 2025. DOI: 10.48550/arXiv.2502.14145.

[6] ***Lewis P, Perez E, Piktus A.*** Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2021. DOI: 10.48550/arXiv.2502.14145.

[7] ***Sia S, Duh K.*** In-context Learning as Maintaining Coherency: A Study of On-the-fly Machine Translation Using Large Language Models. 2023. DOI: 10.48550/arXiv.2305.03573.

[8] ***Arslan M, Ghanem H, Munawar S.*** A Survey on RAG with LLMs // *Procedia Computer Science*. 2024; 246: 3781–3790. DOI: 10.1016/j.procs.2024.09.178.

[9] ***Guu K, Lee K.*** REALM: Retrieval-augmented language model pre-training. 2020. DOI: 10.48550/arXiv.2002.08909.

[10] ***Liu R, Jia Z, Bao F.*** Retrieval-Augmented Dialogue Knowledge Aggregation for expressive conversational speech synthesis // *Information Fusion*. 2025; 118: 102948. DOI: 10.48550/arXiv.2501.06467.

[11] ***Peng B, Zhu Y, Liu Y.*** Graph retrieval-augmented generation: a survey. 2024. DOI: 10.48550/arXiv.2408.08921.

[12] ***Yasunaga M, Ren H, Bosselut A.*** QA-GNN: Reasoning with language models and knowledge graphs for question answering. 2022. DOI: 10.48550/arXiv.2104.06378.

[13] ***Taunk D, Khanna L, Kandru P.*** GrapeQA: GRaph augmentation and pruning to enhance question-answering. 2023. DOI: 10.48550/arXiv.2303.12320.

[14] ***Zhang J, Zhang X, Yu J.*** Subgraph retrieval enhanced model for multi-hop knowledge base question answering // Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers). Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.acl-long.396.

[15] ***Mavromatis C, Karypis G.*** GNN-RAG: Graph neural retrieval for large language model reasoning. 2024. DOI: 10.48550/arXiv.2405.20139.

[16] ***Borovik SY.*** Ontologies, Intersubjective Control and V.A. Vittikh Evergetics [In Russian]. *Ontology of designing*. 2020; 10(3): 255–272. DOI: 10.18287/2223-9537-2020-10-3-255-272.

[17] ***Guarino N, Oberle D, Staab S.*** What is an ontology? // Handbook on ontologies. Springer, 2009. P.1–17. DOI: 10.1007/978-3-540-92673-3_0.

[18] ***Lomov PA.*** Formation of the Lexical Module of Applied Ontology for Its Learning [In Russian]. *Ontology of designing*. 2024; 13(4): 520–530. DOI: 10.18287/2223-9537-2023-13-4-520-530.

[19] ***Lomov PA.*** Automation of Synthesis of Composite Ontological Content Patterns [In Russian]. *Ontology of designing*. 2016; 20(6): 162–172. DOI: 10.18287/2223-9537-2016-6-2-162-172.

[20] ***Lomov PA, Shishaev MG***. Formation of Cognitive Frames Based on Ontological Patterns for Ontology Visualization [In Russian]. *Information Systems and Technologies*. 2015; 6: 12–22.

[21] ***Reimers N, Gurevych I.*** Sentence-BERT: Sentence embeddings using siamese BERT-networks. 2019. DOI: 10.48550/arXiv.1908.10084.

[22] ***Peters ME, Neumann M, Iyyer M.*** Deep contextualized word representations. 2018. DOI: 10.48550/arXiv.1802.05365.

[23] ***Zmitrovich D, Abramov A, Kalmykov A.*** A family of pretrained transformer language models for Russian. 2023. DOI: 10.48550/arXiv.2309.10931.

About the author

Pavel Andreevich Lomov (b.1984), PhD, a senior researcher at the Institute for Informatics and Mathematical Modeling named after V.A. Putilov – Subdivision of Kola Science Centre of the Russian Academy of Sciences. Research interests: knowledge representation, ontological modeling, semantic web. AuthorID (RSCI): 8479-8320. Author ID (Scopus): 55350587100; ORCID: 0000-0002-0924-0188; Researcher ID (WoS): P-6627-2015. *palandlom@yandex.ru.*

Received February 2, 2025, Revised March 10, 2025. Accepted March 20, 2025.