

Предметно-ориентированные языковые модели непрерывного обучения

О. Н. Злобин

*Санкт-Петербургский
государственный
университет
телекоммуникаций
им. проф. М. А. Бонч-Бруевича*
olegzlobin00@mail.ru

В. Л. Литвинов

*Санкт-Петербургский
государственный
электротехнический
университет «ЛЭТИ»
им. В.И. Ульянова (Ленина)*
vlad.litvinov61@gmail.com

Ф. В. Филиппов

*Санкт-Петербургский
государственный
университет
телекоммуникаций
им. проф. М. А. Бонч-Бруевича*
9000096@mail.ru

Аннотация. В работе исследуются вопросы путей создания языковых моделей со встроенным механизмом сознания в предметной области. Предлагается, наряду с вербальной памятью, введение дополнительной когнитивной памяти с доменным администрированием для обеспечения непрерывности обучения. Описывается способ организации сопряжения вербальной и когнитивной памяти. Предложена концептуальная схема реализации искусственного интеллекта в ограниченной предметной области с использованием предметно-ориентированной языковой модели. Обсуждается вопрос смены режимов инференса и обучения.

Ключевые слова: искусственный интеллект, языковые модели, вербальная память, когнитивная память, контекст обновления, эмбединги, домены, концепты онтологий

I. ВВЕДЕНИЕ

Технология больших языковых моделей заключается в обработке входных и генерации выходных данных на уровне анализа токенов. Основным механизмом, который используется для этих целей, является механизм «внимания» [1], а основным рабочим инструментом служит обученная вербальная память [2], позволяющая оценивать наиболее вероятные последовательности слов в процессе их генерации и максимизировать вероятность следующего слова, с учётом предыдущего контекста. Несмотря на колоссальные объёмы вербальной памяти, очевидным «слабым звеном» является завершенность ее обучения. Поэтому обеспечение непрерывности обучения является важной практической задачей.

Очевидно, сложно реализовать искусственный интеллект (ИИ), способный проявлять выдающиеся интеллектуальные способности во всех сферах человеческой деятельности. Точно также как люди являются специалистами в разных областях, естественно ориентировать ИИ на определенную предметную область. Поэтому следует ограничить требования к используемой языковой модели, путем ее ориентации на определенный домен (предметную область). В работе исследуется вопрос построения таких предметно-ориентированных моделей и возможности их непрерывного обучения.

В работе исследуется вопрос построения таких предметно-ориентированных моделей. Ставится задача формирования дополнительной когнитивной памяти для отслеживания и накопления новой информации, а также разработки концептуальной модели ее сопряжения с

предметно-ориентированной языковой моделью для реализации процесса непрерывного обучения.

II. ПРЕДПОСЫЛКИ РЕАЛИЗАЦИИ

Первые попытки приблизить мир компьютеров к «пониманию» контента с целью поиска определенной информации были предприняты путем введения микроформатов и затем микроданных в web-описания. Их назначение – сделать текст для компьютера не просто набором слов, а придать ему семантическое значение. Фундаментальный подход к решению этой задачи был предложен Тимом Бернерс-Ли [3] в виде расширения веб, в котором информация представлена в структурированном и формализованном виде, что позволяет машинам лучше «понимать» и интерпретировать данные. По сути, современный семантический веб, реализованный в стандарте RDF (Resource Description Framework), моделирует отношения между ресурсами и представляет их как связанные структуры.

Модели различных предметных областей на основе онтологий из RDF-графов разрабатываются для того, чтобы информационные системы могли создавать и поддерживать внутренние представления об окружающей среде. Они используются для сбора, обработки и хранения знаний, основанных на семантике слов, что является ключевой задачей в области ИИ. Точно также как человек в своем творчестве обращается к необходимым справочным ресурсам, системы ИИ могут использовать семантически связанные открытые данные – LOD (Linked Open Data) для осознания окружающего мира в любой предметной области [4, 5].

По сути, модель LOD представляет собой постоянно пополняемую глобальную базу знаний в виде агрегации онтологий [6]. Известно, что фундаментальными сущностями онтологий являются концепты, позволяющие формализовать описание объектов любого класса. Использование концептуального подхода позволит смоделировать процесс мышления на чисто семантическом уровне. На практике концепция часто соответствует предложению в текстовом документе или эквивалентному речевому высказыванию. В частности, в фундаментальной работе [7] предложена большая концептуальная модель на базе строителя эмбедингов предложений [8]. При этом, предлагаемая идея не выходит за рамки существующих технологий с использованием одного механизма «внимания», а лишь переводит использование этого механизма на более высокий семантический уровень.

III. ОСНОВНАЯ ИДЕЯ

Вербальная память языковой модели состоит из набора трех матриц W_O , W_K и W_V обученных весовых коэффициентов. Ее обучение прекращается в определенное время и, тем самым, обрывается процесс пополнения ее «знаний» и наступает стадия инференса модели.

Основная идея заключается в том, чтобы наряду с вербальной памятью использовать дополнительную, когнитивную память. Назначение этой дополнительной памяти состоит в постепенном накоплении новой информации, получаемой после прекращения обучения языковой модели. Новые факты можно накапливать в виде URL ссылок на ресурсы LOD и размещать их в когнитивной памяти. Эти ссылки будут служить контекстом обновления. Координатами размещения URL ссылок будет служить пара индексов векторов эмбединга вербальной памяти, которые соответствуют концептам соответствующего факта. Как известно, в процессе администрирования онтологий формируются триплеты, определяющие некоторые новые концепты S_i и S_j через предикаты P их отношений:

$$S_i P \rightarrow S_j.$$

Каждому новому факту, представленному триплетом, будет соответствовать контент, доступный по URL ссылке, которая и будет запоминаться в когнитивной памяти. Поскольку предикаты являются связующим звеном взаимодействующих концептов, когнитивную память можно организовать в виде совокупности N сильно разреженных матриц, где N – число предикатов.

Схема организации сопряжения вербальной и когнитивной памяти представлена на рис. 1. Здесь, в координатах X , Y условно представлена матрица обученных весовых коэффициентов языковой модели – это может быть любая из трех матриц W_O , W_K и W_V , а линия $x_i = y_i$ на ней соответствует идентичным векторам эмбединга. Точки $x_j p_m y_i$ и $x_j p_k y_i$ соответствуют размещению в когнитивной памяти URL ссылок на ресурсы с описанием новых фактов, касающихся векторов x_i и y_i .

Для последовательности входных данных $x \in R^{N \times d}$ механизм внимания с помощью матриц W_O , W_K и $W_V \in R^{d \times d_{in}}$, образующих вербальную память из обученных весовых коэффициентов, вычисляет запросы, ключи и значения:

$$Q = xW_O, \quad K = xW_K, \quad V = xW_V,$$

на базе которых формируются выходные последовательности данных $y \in R^{N \times d}$ с помощью функции softmax:

$$y_i = \sum_{j=1}^i \frac{e^{(q_i^T k_j / \sqrt{d_{in}})}}{\sum_{l=1}^i e^{(q_i^T k_l / \sqrt{d_{in}})}} V_j.$$

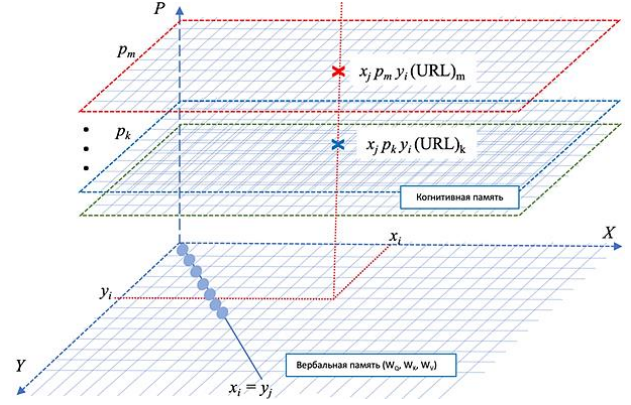


Рис. 1. Схема сопряжения вербальной и когнитивной памяти

Здесь для формирования выхода используется только вербальная память, которая будет формировать предварительный ответ на последовательность входных данных (запрос).

Элементы послыонных разреженных матриц (URL ссылки) будут использоваться как контекст обновления для коррекции выходной последовательности. По сути, подобная коррекция, в терминах механизма внимания может быть представлена как добавочная S_{ij} компонента при формировании выходной последовательности:

$$S_{i,j} = \frac{e^{q_i^T k_j}}{\sum_{l=1}^m e^{q_i^T k_l}} V_j,$$

где q_i и k_j – эмбединг-векторы, соответствующие концептам S_i и S_j нового факта, а V_j – значение, полученное на их основе, m – число концептов, связанных с предикатом P . В этом смысле предикат играет роль дополнительного (нового) ключа, который не участвовал в реализации механизма внимания в обученной модели. Ось P (рис. 1) когнитивной памяти соответствует различным предикатам, которые формируются в процессе администрирования онтологии домена. Хорошо отработанные механизмы сопровождения онтологий являются надежной гарантией поддержания адекватности информации.

IV. СХЕМА РЕАЛИЗАЦИИ

Особый интерес представляет механизм «сознания» ИИ. Трудно говорить о механизме сознания, когда последнее является далеко неопределенным [9]. Поэтому здесь, мы предлагаем придерживаться, по крайней мере, одного из необходимых его свойств – осознания окружающего мира. Таким окружающим миром для ИИ является только та информация, которую предоставляет человек, в том или ином виде.

В этой связи, базовой основой для формирования датасета для обучения предметно ориентированной языковой модели может служить совокупность доменов LOD, соответствующих выбранной предметной области.

При этом, коль скоро речь идет об ИИ, специализирующемся на конкретной предметной области и, возможно некоторых смежных областях, то эта объединенная онтология, заполняемая конкретной информацией, и является его окружающим миром.

Положим, в качестве основы взята некоторая базовая языковая модель, и она прошла обучение на датасете, включающем все сведения из указанной объединенной онтологии. Встает вопрос, как обеспечить непрерывность ее обучения после начала эксплуатации.

Очевидно, что стимулом для активизации механизма «сознания» должен быть факт получения новых данных. При этом может происходить достаточно интенсивное пополнение и модификация информации.

Для обеспечения непрерывности обучения предлагается использовать в качестве промежуточного буфера когнитивную память, описанную выше. Эта память будет абсолютно пустой при первом запуске обученной языковой модели. В процессе эксплуатации модели производится независимое администрирование упомянутой выше объединенной онтологии и все новые факты добавляются в когнитивную память.

Концептуальная схема реализации ИИ представлена на рис. 2. При работе схемы, после получения предварительного ответа на запрос (входной контент) от вербальной памяти, в когнитивной памяти осуществляется поиск наличия новых фактов (URL ссылок), связанных со всеми парами индексов концептов участвующих в предварительном ответе. Если таковых нет (контекст обновления = 0), предварительный ответ считается полным, в противном случае к предварительному ответу добавляется контент обновления и вербальная память запрашивается повторно.

Открытым остается вопрос о том, когда следует выполнять новое обучение вербальной памяти. Очевидно, сигналом к этому может служить состояние когнитивной памяти, например превышение ее допустимого объема или переполнение определенного сегмента. При подобном сигнале осуществляется переход от режима инференса к новому обучению модели. При переключении обратно, в режим инференса, содержимое вербальной памяти сбрасывается.



Рис. 2. Концептуальная схема реализации ИИ

Приведем небольшой пример, иллюстрирующий работу предложенной концептуальной схемы. В качестве предварительно обученной языковой модели возьмем DeepSeek и зададим вопрос: Какие квантовые компьютеры могут работать от обычной розетки?

Получаем ответ: на текущий момент (2024 год) квантовые компьютеры не могут работать от обычной розетки из-за их экстремальных требований к инфраструктуре.

Понятно, что использованная версия DeepSeek обучалась до появления подобных компьютеров. На sparql-запрос о наличии подобных компьютеров мы получаем несколько URL ссылок и, в частности, ссылку <https://equal1.com/news/>. Эта ссылка может служить контекстом обновления. Тогда на повторный запрос: как бы ты ответил на первый вопрос с учетом информации из <https://equal1.com/news/>?

Получаем ответ: компания Equal1 (Ирландия) разработала первый в мире кремниевый квантовый компьютер, который:

- подключается к обычной розетке;

- не требует криогенного охлаждения;
- потребляет мощность, сопоставимую с мощным персональным компьютером (~1–3 кВт).

Конечно, в данном примере активизация механизма сознания проведена в «ручном» режиме. Предложенная концепция позволит реализовать непрерывный процесс обучения. При переходе от режима инференса к дообучению все новые факты будут «усвоены» языковой моделью в вербальной памяти.

V. ЗАКЛЮЧЕНИЕ

В заключение следует отметить, что предлагаемый подход достаточно эффективно решает поставленную задачу обеспечения непрерывности обучения: сочетание использования предметно-ориентированной модели с семантически связанными LOD. Непрерывное накопление новых фактов в когнитивной памяти в виде ссылок позволяет их учитывать в процессе референса, а периодический тюнинг модели и сброс когнитивной памяти обеспечивает непрерывность обучения. По сути, периодический тюнинг – это последовательное дообучение языковой модели.

Полученные результаты можно сравнить с методами генерации с дополненным поиском [10], но, в отличие от этих методов, здесь решается ряд проблем и ограничений, например таких, как точность извлечения данных, потенциально устаревшая информация, несоответствие плотности знаний и избыточное извлечение. Это достигается благодаря сочетанию обширных знаний в ограниченной предметной области с актуальной информацией из внешних источников.

СПИСОК ЛИТЕРАТУРЫ

- [1] Vaswani A., Shazeer N. et al Attention Is All You Need // 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA // URL: <https://arxiv.org/pdf/1706.03762v5> (дата обращения: 10.01.2025).
- [2] Steven Schwering, Maryellen MacDonald. Verbal Working Memory as Emergent from Language Comprehension and Production // URL: <https://pubmed.ncbi.nlm.nih.gov/32226368/>
- [3] Berners-Lee T., Hendler J., Lassila O. The semantic web // Scientific American; May2001, Vol. 284 Issue 5. // URL: http://echo.iat.sfu.ca/library/berners-Lee_01_semantic_web.pdf (дата обращения: 10.01.2025).
- [4] Blaney J. Introduction to the Principles of Linked Open Data // Programming Historian 6 (2017) // URL: <https://doi.org/10.46430/phen0068> (дата обращения: 10.01.2025).
- [5] Губин А.Н., Литвинов В.Л., Филиппов Ф.В., Кравец А.Ю. Визуализация структур данных в результатах запросов к RDF-хранилищам // Вестник СПбГУТ. 2024. Т. 2. № 1.
- [6] The Linked Open Data Cloud URL: <https://lod-cloud.net/> (Accessed 10.01.2025).
- [7] Barrault L., Duquenne P.-A., et al Lage Concept Models: Language Modeling in a Sentence Representation Space // URL: <https://arxiv.org/pdf/2412.08821> (дата обращения: 10.01.2025).
- [8] Duquenne P.-A., Schwenk H., Sagot D. SONAR: sentence-level multimodal and language-agnostic representations, 2023b. // URL: <https://arxiv.org/abs/2308.11466> (дата обращения: 10.01.2025).
- [9] Butlin P., Bengio Y. et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness // arXiv preprint. 2023. URL: <https://arxiv.org/pdf/2308.08708v3.pdf> (дата обращения: 10.01.2025).
- [10] What is Retrieval-Augmented Generation (RAG)? // <https://cloud.google.com/use-cases/retrieval-augmented-generation?hl=ru> (дата обращения: 10.01.2025).