

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ НЕЯВНЫХ ПОЛЬЗОВАТЕЛЬСКИХ ПРЕДПОЧТЕНИЙ В ГЕОИНФОРМАЦИОННЫХ СИСТЕМАХ: НЕЙРОСИМВОЛЬНЫЙ ПОДХОД И ОНТОЛОГИЧЕСКОЕ ДЕКОДИРОВАНИЕ

Е.В. Ледаев¹, А.В. Благов¹

¹Самарский университет, г. Самара
e-mail: egorledyaev1@gmail.com

Аннотация

В работе решается задача извлечения пространственно-временных ограничений из неявных геоинформационных запросов. Предложен нейросимвольный конвейер на базе локальной LLM Qwen 3.5 9B. Внедрение онтологического грамматического декодирования и векторного выравнивания обеспечило 100% валидность на 50 сложных запросах. Учет динамической локации на уровне сэмплера полностью исключает топологические галлюцинации при задержке инференса 4,13 с.

Ключевые слова: нейросимвольный искусственный интеллект, большие языковые модели, геоинформационные системы, парсинг намерений, онтологическое декодирование, пространственный контекст.

Введение

Современные геоинформационные и рекомендательные системы туристической маршрутизации сталкиваются с фундаментальной проблемой «семантического разрыва». Пользователи формулируют намерения на нечетком (fuzzy) естественном языке («тихое место для прогулки с детьми», «болят ноги, хочу кофе с видом»). В то же время алгоритмам маршрутизации на графах (решателям задачи ориентирования с временными окнами - OPTW) и пространственным базам данных (PostgreSQL/PostGIS) требуются строгие, скалярные параметры: бюджет времени, координаты и точные идентификаторы категорий объектов [1].

Традиционные методы извлечения ключевых слов не способны осуществлять контекстный вывод. Использование Больших Языковых Моделей (LLM) решает проблему понимания контекста, однако вероятностная природа авторегрессионных сетей порождает проблему семантических и топологических галлюцинаций [2]. Модель может синтаксически верно извлечь категорию «океанариум» или «метро», отсутствие которых в текущей локации пользователя (например, в малом городе) приведет к краху последующего SQL-запроса.

Целью данной работы является разработка отказоустойчивой гибридной архитектуры извлечения намерений (Intent Encoder) в парадигме нейросимвольного искусственного интеллекта (Neuro-Symbolic AI), объединяющей вероятностную мощь локальных LLM с жесткими детерминированными правилами пространственных онтологий [3].

1. Нейросимвольная архитектура и онтологическое декодирование

Разработанный модуль принимает на вход текст запроса X и геопространственные координаты старта L_{start} . Задача системы – сформировать JSON-структуру Y , абсолютно валидную как синтаксически, так и топологически. В качестве ядра используется квантованная локальная языковая модель Qwen 3.5 9B.

Для устранения топологических галлюцинаций предложен метод онтологического грамматического декодирования (Ontology-Aware Grammar Decoding). В отличие от статических JSON-схем, предложенная архитектура формирует пространство допустимых ответов динамически, на основе символьной логики геоинформационной БД.

Процесс состоит из двух этапов:

1. Пространственное ограничение (Spatial Grounding): На основе L_{start} система выполняет быстрый SQL-запрос ST_DWithin (PostGIS), формируя локальную онтологию O_{loc} – множество категорий объектов (POI), физически присутствующих в заданном радиусе доступности (например, 5 км).

2. Управляемая генерация (Constrained Generation): Множество O_{loc} внедряется в качестве перечислимого типа (enum) в динамическую JSON-схему $S(O_{loc})$.

Генерация ответа Y математически ограничивается схемой $S(O_{loc})$, что позволяет на уровне сэмплера отсекалть вероятности токенов, ведущих к генерации объектов, отсутствующих в локации:

$$P(y_t | y_{<t}, X, O_{loc}) = 0, \text{ если токен } y_t \notin S(O_{loc}). \quad (1)$$

Таким образом, если пользователь запрашивает «пляж» в локации, где данная категория отсутствует в O_{loc} , LLM физически лишена возможности сгенерировать этот токен и вынуждена осуществлять семантический поиск ближайшей доступной альтернативы (например, «набережная» или «парк») из предложенного динамического словаря.

2. Механизм векторного семантического выравнивания (Self-Correction)

Для обработки сложных пограничных случаев (Edge Cases), когда прямое онтологическое декодирование приводит к снижению уверенности модели (Logprobs), в конвейер внедрена асинхронная петля самокоррекции (Feedback Loop) с использованием векторных представлений.

Если модель затрудняется напрямую спроецировать неявное пожелание пользователя на элементы множества O_{loc} , система применяет алгоритм семантического выравнивания (Semantic Alignment). Запрос пользователя или сгенерированная абстрактная концепция c_{gen} преобразуется в вектор $E(c_{gen})$ с помощью эмбединг-модели multilingual-e5-base. Наиболее релевантный аналог $\hat{c}$ из локальной онтологии O_{loc} находится путем максимизации косинусного сходства:

$$\hat{c} = \arg \max_{c \in O_{loc}} \frac{E(c_{gen}) \cdot E(c)}{|E(c_{gen})| |E(c)|}. \quad (2)$$

Если максимальное косинусное сходство оказывается ниже эмпирического порога $\tau = 0.75$, абстрактная категория отбрасывается во избежание смыслового искажения маршрута. Эта двухуровневая защита (символьная на уровне декодера и векторная на уровне бизнес-логики) формирует надежный нейросимвольный пайплайн.

3. Механизм векторного семантического выравнивания (Self-Correction)

Для валидации метода был сформирован набор данных из 50 имплицитных запросов на русском языке (например: «Я с детской коляской, у нас часа три, не хотим идти в музей»). В качестве стартовых локаций были заданы координаты различных городов, обладающих принципиально разной топологической онтологией (от мегаполисов до малых городов).

Для доказательства эффективности нейросимвольного подхода был проведен сравнительный анализ (Ablation Study) с базовыми архитектурами извлечения информации:

- Zero-shot Prompting: базовая генерация без ограничений.
- Static Grammar Decoding: ограничение статической JSON-схемой (содержащей все категории мира, без привязки к локации).
- Ontology-Aware Grammar Decoding (Предложенный метод): генерация с динамической подменой схемы на основе PostGIS-запроса и петель самокоррекции (Формула 2).

Оценка производилась по синтаксической валидности (Raw JSON Success), предметно-пространственной точности без галлюцинаций (Spatial Validity) и среднему времени ответа (Latency). Инференс осуществлялся на графическом ускорителе NVIDIA RTX 2060 SUPER (модель квантована до 4 бит).

Таблица 1 – Сравнительный анализ архитектур извлечения намерений

Архитектура пайплайна	Raw JSON Success, %	Spatial Validity, %	Latency, с
Zero-shot Prompting (Базовая LLM)	78,0	46,0	1,82
Static Grammar Decoding	100,0	64,0	2,90
Ontology-Aware Decoding (Предложенная)	100,0	100,0	4,13

Как видно из таблицы 1, стандартная генерация (Zero-shot) демонстрирует низкий уровень синтаксического качества (22% ошибок парсинга) и фатальную склонность к топологическим галлюцинациям (54% категорий отсутствовали в радиусе поиска). Внедрение статической грамматики (Static Grammar) решило проблему JSON-синтаксиса, однако пространственная валидность составила лишь 64%, так как модель продолжала выбирать релевантные контексту, но физически отсутствующие в городе объекты.

Предложенная нейросимвольная архитектура (Ontology-Aware Decoding) обеспечила абсолютную (100%) структурную и пространственную целостность выходных данных. Увеличение среднего времени инференса до 4,13 с обусловлено накладными расходами на предварительный SQL-запрос (ST_DWithin) для формирования локальной онтологии, перекомпиляцию JSON-схемы в реальном времени и применение векторного выравнивания (уравнение 2). Тем не менее, для асинхронных геоинформационных систем задержка в ~4 секунды является приемлемой и полностью компенсируется нулевым процентом отказов на этапе графовой маршрутизации (OPTW).

Заключение

Проведенное исследование доказывает перспективность применения нейросимвольного искусственного интеллекта в задачах туристической маршрутизации. Объединение детерминированных пространственных ограничений баз данных с вероятностными алгоритмами локальных LLM (Ontology-Aware Grammar Decoding) позволяет на уровне сэмплера пресекать топологические галлюцинации. Предложенный гибридный пайплайн обеспечивает 100% валидность извлекаемых намерений при полном сохранении приватности пользовательских данных. Дальнейшие работы будут направлены на оптимизацию скорости динамической компиляции грамматик и внедрение JIT-ускорения для снижения общей задержки системы.

Список использованных источников

1. Gunawan, A. Orienteering Problem: A survey of recent variants, solution approaches and applications / A. Gunawan, H. C. Lau, P. Vansteenwegen // *European Journal of Operational Research*. – 2016. – V. 255, № 2. – P. 315–332.
2. Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions / L. Huang [et al.] // *arXiv preprint arXiv:2311.05232*. – 2023. – P. 1–35.
3. Kautz, H. The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture / H. Kautz // *AI Magazine*. – 2022. – V. 43, № 1. – P. 105–125.

INTELLIGENT ANALYSIS OF IMPLICIT USER PREFERENCES IN GEOINFORMATION SYSTEMS: NEURO-SYMBOLIC APPROACH AND ONTOLOGY-AWARE DECODING

E.V. Ledyayev¹, A.V. Blagov¹

*¹Samara University, Samara
e-mail: egorledyaev1@gmail.com*

Abstract

The paper solves the problem of extracting spatial and temporal constraints from implicit geoinformation queries. A neural symbolic pipeline based on the local LLM Qwen 3.5 9B is proposed. The implementation of ontological grammatical decoding and vector alignment ensured 100% validity on 50 complex queries. Taking into account the dynamic location at the sampler level completely eliminates topological hallucinations with an inference delay of 4.13 s.

Keywords: neuro-symbolic AI, large language models, geoinformation systems, intent parsing, ontology-aware decoding, spatial context.