DOI 10.37539/2949-1991.2025.35.12.016

Кондаков Сергей Сергеевич, магистрант,
АНО ВО «Российский новый университет»

МЕТОДОЛОГИЯ СОЗДАНИЯ СПЕЦИАЛИЗИРОВАННОГО TEXT-TO-SQL ДАТАСЕТА ДЛЯ ЯЗЫКА ЗАПРОСОВ 1С: ПРЕДПРИЯТИЕ ПРИ ОГРАНИЧЕННОМ КОНТЕКСТЕ

Аннотация. В работе рассматривается задача Text-to-SQL для предметно-ориентированного языке запросов платформы «1С:Предприятие8». Существующие модели (SOTA), обученные на универсальных бенчмарках, демонстрируют низкую эффективность на проприетарных диалектах SQL. Авторами предлагается методология создания компактного обучающего и проверочного датасета, основанная на ограничении пространства поиска.

Ключевые слова: Text-to-SQL, 1С:Предприятие, специализированный датасет, пространство поиска, Schema Pruning, Vocabulary Restriction.

1. Введение

Text-to-SQL – ключевая задача в области NLP для бизнес-аналитики. Современные LLM показывают впечатляющие результаты на стандартных диалектах, что подтверждается лидерами бенчмарков Spider [1] и BIRD [2].

В то же время в корпоративном секторе РФ одним из стандартов де-факто является язык запросов платформы «1С:Предприятие 8», который, хотя и основан на стандартном SQL, обладает уникальными особенностями [3], затрудняющими работу универсальных моделей:

1. Двуязычный синтаксис: доступен английский и русский, однако в прикладных решениях для РФ принято использование русских ключевых слов.
2. Разыменование ссылочных полей: заменяет явное левое соединение таблиц неявным (например, Продажи.Номенклатура.Упаковка.Вес) с использованием точечной нотации.
3. Виртуальные таблицы: параметризуемые сущности, не имеющие прямых аналогов в ANSI SQL.
4. Специфические конструкции: например, сравнение с пустой ссылкой ЗНАЧЕНИЕ (Справочник.Контрагенты.ПустаяСсылка).

Прямое применение (zero-shot prompting) универсальных LLM приводит к галлюцинациям. В работе предлагается Data-Centric AI подход: фокус на качестве данных через создание специализированного датасета с ограниченным контекстом схемы и редуцированным словарем языка запросов «1С:Предприятие 8».

2. Методология

Подход базируется на гипотезе, что для практического применения языковых моделей не требуется полное покрытие языка запросов и схемы данных.

2.1. Ограничение схемы данных (Schema Pruning)

Таблицы, доступные в языке запросов «1С:Предприятие 8», объединяются в группы объектов метаданных – ближайший аналог схемы базы данных ANSI SQL. Каждая группа имеет свои особенности, которые необходимо учитывать при создании запроса (например, для группы РегистрыСведений: виртуальные таблицы СрезПервых, СрезПоследних и поля Период, Регистратор).

В рамках одного домена используется ограниченный набор групп объектов метаданных, поэтому для тестирования или файн-тюнинга языковых моделей имеет смысл применять Schema Pruning – фокусироваться только на релевантных группах, исключая редко используемые.

2.2. Редукция словаря (Vocabulary Restriction)

Для повышения надежности генерации фиксируется подмножество ключевых слов и конструкций. Исключаются избыточные элементы, не востребованные в текущем домене. Сужение пространства поиска минимизирует вероятность галлюцинаций и уменьшает размер обучающего датасета, позволяя применять легковесные модели.

3. Построение Датасета

Процесс формирования датасета комбинирует статический анализ кода и анализ логов, что позволяет создать набор данных даже при отсутствии накопленной статистики.

3.1. Статический анализ

Основным источником для формирования ограниченной схемы и словаря выступает исходный код прикладных решений «1С:Предприятие 8». Из него извлекаются тексты запросов, из которых отбираются наиболее частотные таблицы и конструкции языка.

3.2. Анализ логов (Log Mining)

В качестве дополнительного источника данных используются записи Технологического журнала «1С:Предприятие 8». Этот подход позволяет собрать реальную статистику по применению языка запросов.

3.3. Синтетическая генерация (Synthetic Pair Generation)

Данные, извлеченные из кода и логов (пп. 3.1, 3.2), содержат только результирующие запросы без исходных вопросов пользователей. Для получения пар «вопрос-ответ» применяется метод обратной генерации (Reverse Engineering via LLM) – мощная модель-учитель, анализируя текст запроса и контекст схемы данных, синтезирует эквивалентный вопрос на естественном языке.

4. Обсуждение результатов

Статистика упоминания групп таблиц в запросах прикладного решения 1С:ERP 2.5 (таблица 1) подтверждает целесообразность Schema Pruning. Доля некоторых групп достаточно низкая: суммарный вклад, начиная с РегистрБухгалтерии, всего 1,54%.

Таблица 1.

Доля групп таблиц в 1С:ERP 2.5

Группа таблиц	Количество упоминаний	Доля, %
Перечисление	84 102	36,34
Справочник	59 831	25,85
Документ	40 433	17,47
РегистрСведений	20 079	8,68
РегистрНакопления	9 299	4,02
ПланСчетов	8 018	3,46
ПланВидовХарактеристик	6 332	2,74
РегистрБухгалтерии	1 652	0,71
ПланВидовРасчета	669	0,29
Задача	276	0,12
Константа	260	0,11
ПланОбмена	223	0,10
РегистрРасчета	103	0,04
БизнесПроцесс	85	0,04

ЖурналДокументов	41	0,02
КритерийОтбора	36	0,02

Аналогично, статистика использования ключевых слов 1С:ERP 2.5 (таблица 2) подтверждает возможность Vocabulary Restriction.

Таблица 2.

Доля ключевых слов запросов 1С:ERP 2.5

Ключевое слово	**Количество упоминаний**	**Доля, %**
КАК	827 992	39,07
И	166 013	7,83
ЗНАЧЕНИЕ	137 396	6,48
ЕСТЬNULL	112 966	5,33
ВЫБРАТЬ	101 770	4,80
ИЗ	98 230	4,63
ВЫБОР	70 280	3,32
СОЕДИНЕНИЕ	60 912	2,87
ГДЕ	59 850	2,82
НЕОПРЕДЕЛЕНО	39 808	1,88
В	38 648	1,82
НЕ	37 896	1,79
СУММА	37 040	1,75
ПОМЕСТИТЬ	31 243	1,47
ИЛИ	27 708	1,31
ЛОЖЬ	21 368	1,01
NULL	18 475	0,87
ИСТИНА	17 782	0,84
ОБЪЕДИНИТЬ	17 761	0,84
Прочие слова	196 278	9,26

Для экосистемы «1С:Предприятие 8» переход от универсальных Text-to-SQL датасетов к узкоспециализированным – перспективный путь внедрения генеративного ИИ.

Список литературы:

1. Yu, T., et al. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. EMNLP, 2018.

2. Li, J., et al. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs. arXiv, 2023.

3. Хрусталева Е. Ю. Язык запросов «1С:Предприятие 8». Издание 3, стереотипное. М.: ООО «1С-Паблишинг», 2025.

4. Lei, W., et al. Re-examining the Role of Schema Linking in Text-to-SQL. 2020.