

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ПОДДЕРЖКИ ОБУЧЕНИЯ СПЕЦИАЛИСТОВ РАДИОРЕЛЕЙНОЙ СВЯЗИ НА ОСНОВЕ АРХИТЕКТУРЫ RAG

Савицкий А.Ю.¹, Бардашевич А.В.²

¹Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», г. Минск, Республика Беларусь, a.savitskij@bsuir.by

²Воинская часть Вооруженных Сил Республики Беларусь, г. Минск, Беларусь, sasha.bard9364@gmail.com

Abstract. The article considers the problem of improving the quality of training of radio-relay communication specialists through the introduction of an intelligent assistant based on the RAG (Retrieval-Augmented Generation) architecture. The developed system «AXON» provides semantic search and generation of reliable answers only on the basis of a verified knowledge base, operates in a local secure loop without access to external networks. The structure, algorithms and test results of the system are described. The application of RAG technology made it possible to combine the flexibility of a dialogue interface with strict requirements for reliability and information security in military education.

Введение.

В современном мире высшее образование подготовка специалистов радиорелейной связи (РРС) является одной из ключевых задач военного образования, поскольку от уровня их квалификации зависит устойчивость управления войсками. Современные радиорелейные станции представляют собой сложные программно-аппаратные комплексы, требующие глубокого знания технической документации (ТД), инструкций по эксплуатации и регламентов обслуживания. Традиционные методы обучения, основанные на статичных учебных материалах (бумажных носителях, PDF-файлах), не обеспечивают оперативного доступа к информации в условиях ограниченного времени и частого обновления парка техники [1].

Развитие технологий искусственного интеллекта открывает новые возможности для создания интеллектуальных систем поддержки учебного процесса. Однако применение открытых языковых моделей общего назначения недопустимо в военной сфере из-за рисков утечки конфиденциальных данных и генерации недостоверных ответов. Решением является архитектура RAG (Retrieval-Augmented Generation) – гибридный подход, сочетающий семантический поиск по верифицированным источникам с генерацией ответов на основе найденного контекста [2].

RAG позволяет создавать защищённые локальные системы, работающие исключительно с внутренней базой знаний (БЗ) без обращения к внешним серверам.

Цель данной работы – разработка защищённого электронного ресурса «АКСОН» для обучения специалистов РРС, обеспечивающего достоверную генерацию ответов на основе верифицированной БЗ в локальном контуре. В статье представлены архитектура, алгоритмы и результаты испытаний разработанной системы.

1. Анализ предметной области и выбор архитектуры

Обучение специалистов РРС имеет ряд особенностей: большой объём нормативной документации, высокие требования к точности знаний, необходимость оперативного поиска информации при нештатных ситуациях. Существующие системы управления обучением, тренажёры и электронные библиотеки не в полной мере решают эти задачи

из-за статичности контента, сложности навигации и отсутствия интеллектуального поиска по смыслу [1].

Архитектура RAG, предложенная в 2020 г. [2], позволяет преодолеть ограничения традиционных систем. Её ключевые преимущества: достоверность ответов (генерация только на основе предоставленного контекста), верифицируемость (цитирование источников), информационная безопасность (полностью локальное развертывание), оперативная актуализация БЗ (достаточно переиндексировать новые документы без переобучения модели). Эти свойства делают RAG идеальной основой для создания обучающих систем военного назначения.

На рисунке 1 представлена схема работы RAG в контексте обучения специалистов РРС. Пользовательский запрос преобразуется в векторное представление, затем выполняется семантический поиск в БЗ, и на основе найденных фрагментов языковая модель генерирует ответ с указанием источников.



Рисунок 1 – Концептуальная схема RAG

2. Структурное и функциональное проектирование системы «АКСОН»

Система «АКСОН» построена по клиент-серверной архитектуре с выделением следующих функциональных блоков (рис. 2):

- блок пользовательского интерфейса;
- блок аутентификации и контроля доступа;
- блок управления БЗ (загрузка/удаление документов в форматах PDF, DOCX);
- блок предварительной обработки и векторизации (очистка, сегментация, генерация эмбеддингов);
- блок семантического поиска (векторная БД на базе PostgreSQL, индекс HNSW, косинусное расстояние, порог релевантности 0,55);
- блок формирования контекста (сборка промпта);
- блок генерации ответов (локальная языковая модель Qwen 2.5:0.5b на сервере Ollama);
- блок кэширования результатов;
- блок логирования и обработки ошибок.

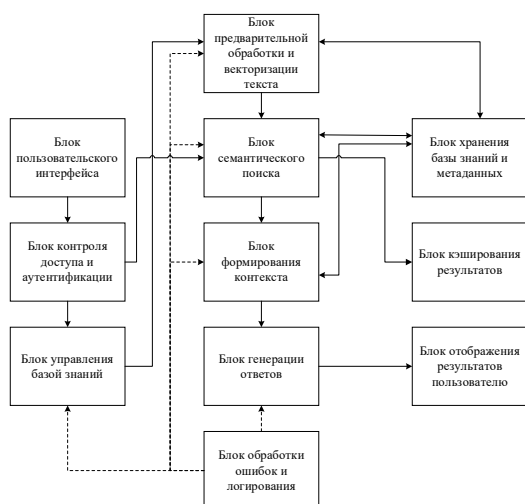


Рисунок 2 – Структурная схема системы

Функциональная схема обработки запроса включает следующие этапы:

Ввод запроса пользователем, аутентификация по JWT-токену.

Проверка кэша (по SHA-256 хешу запроса) – при наличии ответ возвращается мгновенно.

Векторизация запроса и семантический поиск в БД (топ-3 наиболее релевантных чанка).

Проверка максимальной релевантности: если $< 0,55$ – формируется ответ «Информация не найдена».

Очистка чанков от артефактов (URL, ссылки, номера страниц), объединение в контекст, добавление системного промпта.

Потоковая генерация ответа языковой моделью с параметрами: температура 0,1, max_tokens 800.

Добавление блока источников (названия документов и проценты релевантности).

3. Модель данных и ключевые алгоритмы.

Модель данных включает пять основных таблиц: users (логин, хеш пароля, роль, активность), documents (имя файла, путь, хеш, статус обработки, загрузивший пользователь), knowledge_chunks (текст чанка, эмбединг vector(384), метаданные JSONB, ссылка на документ), chats (название, владелец), messages (роль, текст, источники JSONB, временная метка). Для ускорения поиска созданы индексы, в том числе HNSW на векторном поле.

Алгоритм семантического поиска использует оператор косинусного расстояния ($\leq$) расширения pgvector, фильтрацию по порогу 0,55 и лимит 3 результата. При генерации ответа система строго следует инструкции: «Используй только предоставленный контекст. Если в контексте нет информации – скажи об этом. Укажи источники».

4. Реализация и тестирование.

Программная реализация выполнена на Python (FastAPI) для серверной части и нативном JavaScript для клиентской. Все компоненты развернуты локально в изолированном контуре.

Тестирование (табл. 1) проводилось на оборудовании: Intel Core i5-1135G7, 16 ГБ ОЗУ, SSD, ОС Windows 11. Проверены функциональные сценарии: аутентификация, управление чатами, загрузка

и удаление документов, семантический поиск, генерация ответов с источниками, административные функции.

Таблица 1 – Результаты функционального тестирования (фрагмент)

Проверка	Результат
Вход с корректными данными	Успешно
Загрузка PDF/DOCX, индексация	Успешно (среднее время 3,2 с для файла 5 МБ)
Поиск релевантного запроса	Топ-3 чанка, релевантность $>0,55$
Генерация ответа с источниками	Ответ содержит блок источников
Отсутствие информации в БЗ	Сообщение «Информация не найдена»
Изменение роли пользователя (админ)	Успешно

Временные характеристики соответствуют нормативам: среднее время генерации ответа – 3,5 с (при штатной нагрузке), что не превышает 7 с, установленных в техническом задании.

5. Преимущества для образовательного процесса.

Разработанная система «АКСОН» позволяет:

сократить время поиска информации по ТД с нескольких минут до нескольких секунд;

получать ответы на естественном языке, что снижает когнитивную нагрузку на обучаемых;

обеспечивать достоверность ответов за счёт привязки к верифицированным источникам;

оперативно обновлять БЗ при поступлении новых образцов техники без переобучения модели;

проводить обучение в автономном режиме, что критично для полевых условий и учебных классов без выхода в интернет.

Таким образом, использование RAG-технологии в военном образовании позволяет повысить качество подготовки специалистов, сохраняя при этом строгие требования информационной безопасности.

В работе представлена разработка интеллектуальной системы «АКСОН» для поддержки обучения специалистов радиорелейной связи на основе архитектуры RAG. Система обеспечивает семантический поиск и генерацию достоверных ответов в локальном защищённом контуре. Проведённые испытания подтвердили выполнение функциональных и временных требований. Разработанный ресурс готов к внедрению в учебный процесс военного факультета и может быть рекомендован для использования в подразделениях связи Вооружённых Сил Республики Беларусь. Дальнейшее развитие видится в расширении БЗ, интеграции с другими языковыми моделями и добавлении мультимодальных возможностей.

Литература

1. Утин, Л.Л. Военные системы PPC: учеб.-метод. пособие / Л.Л. Утин, В.А. Федоренко, Е.А. Масейчик. – Минск : БГУИР, 2020. – 158 с.

2. Lewis, P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis, E. Perez, A. Piktus et al. – arXiv:2005.11401, 2020.