

АВТОМАТИЧЕСКАЯ СИСТЕМА ГЕНЕРАЦИИ ТЕМАТИЧЕСКИХ ДИАЛОГОВ И ПЕРСОНАЛИЗИРОВАННЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ RAG-АРХИТЕКТУРЫ ДЛЯ ИЗУЧЕНИЯ ИНОСТРАННЫХ ЯЗЫКОВ

Пинчукова П.В. (ИТМО), Миронова А.А. (ИТМО), Молокова М.А. (ИТМО)

Научный руководитель — кандидат технических наук, доцент Федоров Д.А. (ИТМО)

Введение. В настоящее время умение говорить на иностранных языках становится ключевым фактором для профессионального роста и личностного развития. Однако стандартные методы заучивания слов часто не учитывают индивидуальные особенности учащихся, предлагая однотипные упражнения и обобщенные тексты по тематике. Это значительно снижает эффективность запоминания и интерес к изучению.

Для решения этой задачи предлагается разработать автоматическую систему (в перспективе — полноценное приложение), которая будет предлагать тексты, максимально соответствующие интересам обучающегося, а также использовать лексику, введенную самим пользователем. Система также должна обладать функцией диалога, учитывая культурный контекст и ситуационные особенности. Проект актуален, поскольку современные исследования в области когнитивной лингвистики подтверждают, что слова и выражения иностранного языка лучше всего запоминаются и усваиваются, когда они находятся в правильном контексте, а не изолированы от него. Кроме того, растет спрос на персонализированное обучение, а технологии обработки естественного языка (NLP) развиваются, позволяя автоматически создавать учебные материалы. Основная цель разработки — создание продукта, который будет служить помощником в запоминании и практике языка.

Основная часть. Система представляет собой автоматизированную платформу для создания текстов и диалогов, направленных на изучение иностранных языков. Она работает на основе архитектуры Retrieval-Augmented Generation (RAG), которая разработана для расширения возможностей больших языковых моделей благодаря интеграции мощных механизмов поиска информации (компоненты поиска — retriever и генерации — generator).

В контексте изучения иностранных языков будет реализована генерация диалогов и текстов путем извлечения релевантных примеров из корпусов (грамматика, идиомы, культурные сценарии), адаптированных под уровень ученика (A1–C2) и цели. Первый сценарий использования заключается в том, что пользователи загружают список слов для изучения, выбирают интересующие их темы и указывают уровень владения языком. Эта информация преобразуется в векторные представления, проводится семантический поиск релевантного контекста, после чего формируется связный текст с использованием заданной лексики и переводом. Второй сценарий позволяет вводить контекст и желаемые слова, после чего генерируется скрипт краткого диалога, максимально приближенного к реальной жизни. Применение RAG-архитектуры гарантирует точность по тематике и снижает риск генерации некорректных данных. Персонализация содержания и релевантный контекст способствует более глубокому усвоению материала.

Выводы. В результате будет разработан продукт, основанный на технологиях обработки естественного языка (NLP), машинного обучения (ML) и архитектуре RAG. Он обеспечит создание интерактивной обучающей среды, которая станет альтернативой традиционным методам заучивания.

Список использованных источников:

1. Боярский К. К. Введение в компьютерную лингвистику: учебное пособие. – СПб: НИУ ИТМО, 2013. – 72 с.
2. Коган М. С., Попова Н. В. Использование корпусных технологий в обучении иноязычной лексике // Известия Российского государственного педагогического университета им. А. И. Герцена. —2023. — № 207. — С. 145–155.
3. Оболенский Д.М., Шевченко В.И. 2024. Использование метода RAG и больших языковых моделей в интеллектуальных образовательных экосистемах. Экономика. Информатика, 51(3): 699–709. DOI 10.52575/2687-0932-2024-51-3-699-709
4. Retrieval-augmented generation for educational application: A systematic survey — Li Zongxi, Zijian Wang, Weiming Wang — «Computers and Education: Artificial Intelligence», 2025