А. А. Шеремет, И. А. Устинов, Ю. А. Плотоненко
Тюменский государственный университет, г. Тюмень
УДК 004.891

РАЗРАБОТКА ЧАТ-БОТА ДЛЯ ФОРМИРОВАНИЯ ИНДИВИДУАЛЬНЫХ ИНСТРУКЦИЙ (ДОРОЖНЫХ КАРТ) С ИСПОЛЬЗОВАНИЕМ ИНСТРУМЕНТОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ КЛИЕНТОВ МФЦ ТЮМЕНСКОЙ ОБЛАСТИ

Аннотация. Статья посвящена разработке чат-бота для автоматизации создания персонализированных инструкций для МФЦ. Цель — повысить доступность госуслуг за счет сокращения времени поиска информации и минимизации ошибок при сборе документов. Использована гибридная модель (BM25 + BERT) для точности поиска услуг (F1=0.96, 94%) и языковые модели (DeepSeek-V3, Llama-3, Gemma), преобразующие данные в инструкции.

Ключевые слова: чат-бот, NLP, BERT, языковые модели, RAG, МФЦ, государственные услуги.

Введение. Современные многофункциональные центры (МФЦ) сталкиваются с проблемами неэффективного взаимодействия с пользователями, включая необходимость многократных посещений из-за недостаточной информированности о перечне документов и ошибок при их заполнении. Это увеличивает временные затраты граждан и нагрузку на сотрудников.

В ряде исследований подчеркивается важность создания удобных пользовательских интерфейсов для взаимодействия граждан с государственными сервисами. Например, в статье [1] Ивановой А.А. указано, что интеграция государственных услуг позволит исключить лишние запросы для уточнения списка необходимых документов и сократить итоговое количество обращений в МФЦ. Также автор отмечает, что необходим персонализированный подход к предоставлению информации об услугах в интерфейсе.

Цифровизация государственных услуг является приоритетным направлением развития электронного государства. Применение электронных сервисов способствует снижению нагрузки на сотрудников, экономии времени пользователей и оптимизации работы учреждений. В частности, использование технологий искусственного интеллекта открывает новые возможности в автоматизации процессов, предоставлении точных рекомендаций и адаптации информации под конкретные запросы граждан [2].

Так, интеллектуальные системы способны: анализировать пользовательские запросы и формировать точный перечень необходимых документов; проверять корректность заполнения форм; выделять из документации только релевантную информацию; обеспечивать персонализированное взаимодействие.

Также в последние годы наблюдается активное внедрение технологий обработки естественного языка и больших языковых моделей в различные сферы, включая государственные услуги. Одним из перспективных подходов является Retrieval-Augmented Generation (RAG) — архитектура, которая сочетает возможности генеративных моделей с механизмами извлечения информации из внешних источников. RAG позволяет моделям обращаться к актуальным данным вне их обучающего корпуса, обеспечивая более точные и контекстуально релевантные ответы.

Примерами положительного использования данной архитектуры в государственных структурах являются: KemenkeuGPT (Индонезия) — RAG-система для обработки финансовых документов, повысившая точность ответов [3, 4]. LegalRAG (Бангладеш) — многоязычный доступ к правовой информации [5]. Корпоративный опыт — чат-бот Morgan Stanley на базе OpenAI ускорил обработку запросов к 100 000+ документам [6].

Проблема исследования. Анализ текущих процессов взаимодействия граждан с МФЦ выявил следующие проблемы: необходимость многократных визитов из-за отсутствия точной и полной информации о перечне необходимых документов, ошибки при их подготовке, вызванные недостаточной информированностью пользователей.

Несмотря на то, что данные хранятся в базе данных МФЦ, доступ к которой теоретически может быть предоставлен любому пользователю, на практике это требует наличия технических знаний, что создает дополнительный барьер для граждан. Кроме того, документы, предоставляемые МФЦ, часто написаны сложным "канцелярским" языком, который не всегда понятен пользователю и содержит избыточную информацию. В большинстве случаев пользователю важна лишь конкретная часть документа, связанная с его запросом, а не весь объем данных. Поэтому для эффективного решения проблемы необходимо не только упростить доступ к базе данных, но и разработать механизм, позволяющий извлекать из документов только ту информацию, которая непосредственно связана с запросом пользователя.

Исходя из вышесказанного были выделены следующие вопросы, которые определяют функционал системы, решающей данные проблемы:

1. Каким образом можно обеспечить пользователям простой и интуитивно понятный доступ к базе данных МФЦ без необходимости обладания техническими знаниями?

2. Как извлекать из документов только ту информацию, которая непосредственно связана с запросом пользователя, исключая избыточные данные?

3. Какие методы и технологии наиболее эффективны для обработки документов, написанных сложным "канцелярским" языком, и их адаптации для понимания пользователями?

Цель исследования: Разработка чат-бота для МФЦ, генерирующего персонализированные "дорожные карты", подготовки документов с использованием BERT и RAG.

Материалы и методы исследования. В рамках исследования анализировались текста документов МФЦ, содержащие официальные требования к различным услугам. Данные включают текстовые документы, разбитые на секции, содержащие описания услуг. База данных МФЦ — это документоориентированная БД, данные в которой хранятся в иерархической древовидной структуре. Есть корневой документ, в котором хранятся все остальные документы, внутри документа есть секции, в секциях есть подсекции, листом является текст. После сбора данные были агрегированы в один файл, примерная структура файла представлена в табл. 1.

Помимо документов из базы данных также были собраны вопросы, с которыми люди обращаются в МФЦ. Вопросы были собраны путем опроса, проведенного с использованием Яндекс Форм. Участники опроса вводили свои потенциальные запросы, с которыми они бы могли обратиться в МФЦ, после чего сервис, использующий большие языковые модели сопоставил их с названиями соответствующих услуг. Затем, на основе полученной информации, эксперты проводили разметку, определяя, какие секции документов наиболее релевантны для каждого запроса. Дата проведения опроса — февраль 2025 года, количество

респондентов — 19. Возраст респондентов от 18 до 55 лет. В результате было собрано 55 записей, содержащих запросы пользователей.

Таблица 1

Структура собранных данных

Тип услуги	*Название услуги*	*Название секции*	*Текст секции*
Услуги в разрезе органов власти	Осуществление процедуры внесудебного банкротства гражданина	Кто может получить услугу?	"... перечень категорий людей, которые могут получить услугу …"
Жизненные ситуации	Принятие решения о предоставлении льготного проезда	Часто задаваемые вопросы	"... перечень часто задаваемых вопросов и ответы на них…"

Таблица 2

Структура датасета, собранного на основе опроса пользователей

Запрос пользователя	*Секция документа*	*Метка*
Что нужно для загранпаспорта	Кто может получить услугу?	1
Что нужно для загранпаспорта?	Возврат госпошлины	0
Как вернуть подоходный налог?	Обязательные документы	1
Как вернуть подоходный налог?	В электронном виде	0

Данные, использованные в обучении, включают текстовые запросы пользователей и секции документов, которые подходят под эти запросы. Итоговая структура данных состоит из трех ключевых элементов: (1) запрос пользователя, (2) секция документа, (3) метка (релевантность секции). Примерная структура датасета представлена в таблице 2.

Перейдем к решению проблем взаимодействия пользователей с МФЦ через ответы на три исследовательских вопроса

Вопрос 1. Каким образом можно обеспечить пользователям простой и интуитивно понятный доступ к базе данных МФЦ без необходимости обладания техническими знаниями?

Интерфейс: Чтобы позволить пользователю получать нужные данные без обладания технических знаний, необходимо создать интерфейс, который бы позволил пользователю в свободном формате задавать произвольный вопрос, а также получать ответ на него, независимо от местоположения.

Вопрос 2. Как извлекать из документов только ту информацию, которая непосредственно связана с запросом пользователя, исключая избыточные данные?

Поиск релевантных частей. Для решения задачи извлечения из документов только той информации, которая непосредственно связана с запросом пользователя, можно использовать как классические подходы, такие как TF-IDF и BM25, так и современные методы на основе нейронных сетей, такие как BERT.

Подходы, использованные в проекте. В проекте применялся гибридный подход, сочетающий методы ранжирования на основе ключевых слов (BM25) и семантического поиска (BERT). Такой подход позволяет эффективно отбирать релевантные документы, учитывая как точное совпадение терминов, так и смысловую близость запроса и текста.

Предварительный отбор кандидатов с помощью BM25. На первом этапе система использует алгоритм BM25 для быстрого сужения круга потенциально релевантных документов. Этот метод анализирует частоту встречаемости слов запроса в документах, взвешивая их значимость с учетом длины текста. BM25 применяется к текстам и заголовкам документов с последующим комбинированием оценок для повышения точности поиска.

Семантическое переранжирование с помощью BERT. После получения топ-5 кандидатов от BM25 их векторные представления, полученные с помощью предобученной мультиязычной модели DistilUSE (модель на основе архитектуры BERT, предобученная на нескольких языках, включая русский), сравниваются с вектором запроса в семантическом пространстве. Для этого используется библиотека FAISS, обеспечивающая быстрый поиск ближайших соседей по метрике L2 (евклидово расстояние). Переранжирование позволяет скорректировать результаты BM25, отдавая предпочтение документам, которые не обязательно содержат точные формулировки из запроса, но соответствуют ему по смыслу.

Особенности реализации. Использование FAISS для поиска по подмножеству векторов (а не по всему индексу) снижает вычислительные затраты без потери точности. Коэффициенты в формуле комбинирования оценок BM25 (0.7/0.3) могут быть адаптированы под специфику данных на основе анализа качества поиска.

Поиск релевантных частей документов. Документы МФЦ содержат текст, логически разделенный на секции (например, "Кто может получить услугу", "Какие документы нужны" и т. д.). Число секций в каждом документе не одинаково и может варьироваться от 1 до 22 секций. Медианное число секций в документах равняется 9. Зачастую пользователю не интересны все части документа, а важна лишь часть, связанная с его запросом, поэтому для решения этой проблемы необходимо определить, какие секции документа соответствуют конкретному запросу пользователя. Также решение данной задачи позволит уменьшить контекстное окно документа, которые будет подаваться в языковую модель для генерации дорожной карты.

Дано:

- Набор документов $D = \{d_1, d_2, ..., d_n\}$, где d_i — текст документа, n = 1, 2, …, 249.
- Набор секций внутри документа d_i, $S = \{s_{i1}, s_{i2}, ..., s_{im}\}$, где s_{im} — секция внутри документа d_i, m <= 22, медиана — 9 секций.
- Набор пользовательских запросов $Q = \{q_1, q_2, ..., q_k\}$, где q_i — текст запроса пользователя, k = 1, 2, …, 55.
- Разметка данных, содержащая пары (s_{im}, q_i), отмеченные как релевантные (1) или нерелевантные (0).

Задача: Для каждого запроса q_i необходимо определить, какие секции документа s_{im} являются релевантными, т.е. важными для ответа на этот запрос.

Вопрос 3. Какие методы и технологии наиболее эффективны для обработки документов, написанных сложным "канцелярским" языком, и их адаптации для понимания пользователями?

Проблематика. Даже при обеспечении доступа к документам МФЦ, их сложный канцелярский стиль остается барьером для восприятия пользователями. Это снижает эффективность взаимодействия и увеличивает риск ошибок при подготовке документов.

Предлагаемые методы. Большие языковые модели — обеспечивают семантический анализ и перефразирование текста. Retrieval-Augmented Generation (RAG) — комбинирует

извлечение релевантных фрагментов из документов с их последующей адаптацией для пользователя.

Реализация: Пользователь указывает запрос и соответствующую услугу. Система идентифицирует ключевые фрагменты текста из базы данных МФЦ.

Генерация ответа: Извлеченные данные передаются в модель LLM. Модель структурирует информацию в виде пошаговой "дорожной карты".

Минимизация ошибок при подготовке документов. Доказанная эффективность в корпоративных решениях (аналоги: [6]). Данный метод позволяет преодолеть языковой барьер между официальными документами и конечными пользователями, повышая доступность услуг МФЦ.

Результаты. В качестве основной модели была взята DeepPavlov/rubert-base-cased. Для дообучения было взято 5 эпох, важными итоговыми метриками являются: Accuracy = 0.94, F1 = 0.96. В тестовой выборке было 54 секции (20% всего датасета). Языковые модели сравнивались по метрикам: BLEU, ROUGE-N, BERT-Score, среднее время генерации. Результаты тестирования представлены в табл. 3.

Таблица 3

Сравнение языковых моделей

Название модели	*Среднее BLEU (↑)*	*Среднее ROUGE-1 (↑)*	*Среднее ROUGE-2 (↑)*	*Среднее ROUGE-L (↑)*	*Среднее BERT-score (↑)*	*Среднее время генерации, сек. (↓)*	*Кол-во тестов*
gemma-3-27b	0.0046	0.213	0.102	0.207	0.707	61.9	10
Llama-3.1-8B	**0.0096**	0.230	**0.124**	0.223	0.697	**58.1**	10
Llama-3.3-70B	0.0076	**0.231**	0.120	**0.227**	0.706	62.4	20
Llama-3.1-405B	0.0059	0.216	0.104	0.211	**0.711**	59.9	10
DeepSeek-V3	0.0070	0.223	0.112	0.218	0.709	59.5	20

В период тестирования сервиса пользователи обратились к чат-боту 60 раз. По итогу работы системы были выявлены следующие результаты: в 47 случаях бот успешно сформировал список услуг, подходящих под запрос пользователя, а в 13 случаях запросы остались необработанными из-за отсутствия соответствующих услуг в базе знаний. Ниже представлен скриншот диалога, демонстрирующего результат взаимодействия с чат-ботом в период тестирования (см. рис. 1).

Заключение. Разработанный чат-бот на основе гибридной модели (BM25 + BERT) и языковых моделей успешно решает задачу формирования индивидуальных дорожных карт для клиентов МФЦ Тюменской области, демонстрируя высокую точность (F1 = 0.96) и адаптивность к пользовательским запросам (78.3% успешных обработок). Перспективы проекта включают улучшение поиска внутри базы данных для снижения доли необработанных запросов (21.7%), оптимизацию скорости генерации ответов и интеграцию дополнительных NLP-моделей для улучшения интерпретации сложных запросов.

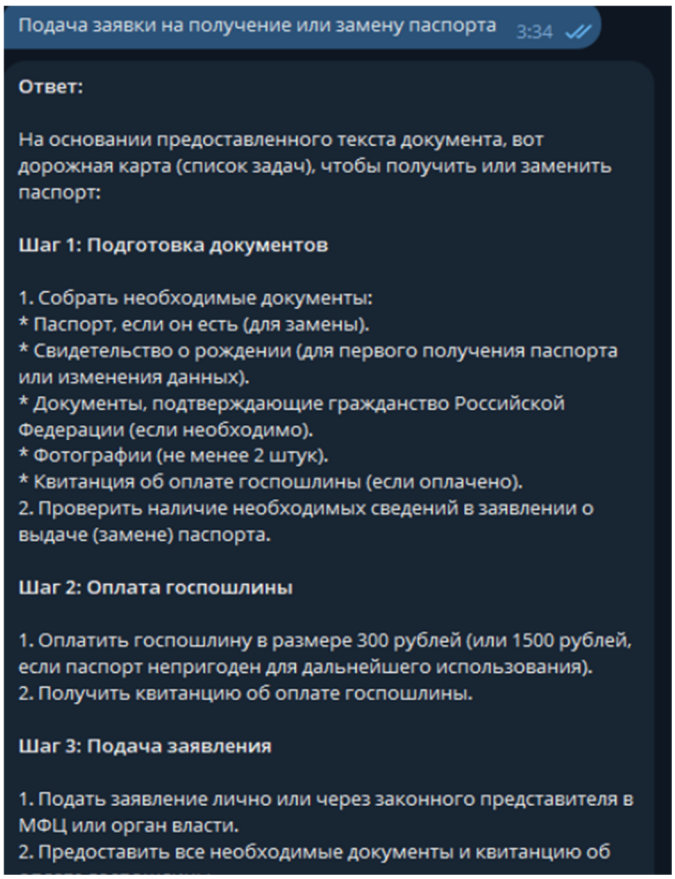

Рис. 1. Ответ чат-бота после выбора услуги

СПИСОК ЛИТЕРАТУРЫ

1. Иванова, А. А. Предоставление "комплексных" государственных и муниципальных услуг по "жизненным ситуациям" / А. А. Иванова // Актуальные проблемы юриспруденции : сборник статей по материалам XIV международной 6научно-практической конференции. Том № 9 (13) : Ассоциация научных сотрудников "Сибирская академическая книга", 2018. — С. 5-9. — EDN XZMFIT.
2. Соловьева, Е. А. Оказание услуг в электронной форме (МФЦ) / Е. А. Соловьева // Трансформация права в информационном обществе : Материалы II Всероссийского научно-практического форума молодых ученых и студентов, Екатеринбург, 21–22 ноября 2019 года. — Екатеринбург: Федеральное государственное бюджетное образовательное учреждение высшего образования "Уральский государственный юридический университет", 2019. — С. 368-374. — EDN NYHVSV.
3. Iqbal, A., Utomo, A. W., Novendra, A. H., Muchtar, A., Syafrullah, M., & Rakhmani, V. (2024). KemenkeuGPT: Leveraging Retrieval-Augmented Generation for Government AI Assistant. arXiv preprint arXiv:2407.21459. URL: https://arxiv.org/abs/2407.21459
4. Шмат, А. В. Применение больших языковых моделей и технологии Retrieval-Augmented Generation для корпоративных ассистентов / А. В. Шмат // Известия Тульского государственного университета. Технические науки. — 2024. — № 10. — С. 720-726. — DOI 10.24412/2071-6168-2024-10-720-721. — EDN GSNKPM.
5. Haque, R., Chowdhury, F. N., Rahman, M. M., Nayeem, M. R., & Zaman, T. (2024). LegalRAG: A Multilingual Retrieval Augmented Generation Based Legal Assistant for Bangladeshi Laws. arXiv preprint arXiv:2504.16121. URL: https://arxiv.org/abs/2504.16121
6. Morgan Stanley.. URL: https://openai.com/customer-stories/morgan-stanley (дата обращения: 20.02.2024).