А. Б. Ибраева, М. С. Цыганова
Тюменский государственный университет, г. Тюмень
УДК 004.912

РАЗРАБОТКА ПОДСИСТЕМЫ АВТОМАТИЧЕСКОГО СТРУКТУРИРОВАНИЯ ИНФОРМАЦИИ ИЗ МЕДИЦИНСКИХ ДОКУМЕНТОВ

Аннотация. В статье представлена разработка подсистемы автоматического структурирования информации из свободного текста, описывающего медицинский осмотр, для сокращения времени, затрачиваемого медицинским персоналом на заполнение электронных форм.

Ключевые слова: автоматическое структурирование текста, обработка естественного языка (NLP), большие языковые модели (LLM), медицинская информационная система (МИС), электронный медицинский документ.

Введение. В связи с активно ведущейся цифровизацией здравоохранения среди многих задач информационного сопровождения лечебно-диагностических процессов возникает необходимость в реализации автоматического структурирования информации из медицинских документов. Согласно приказу Министерства здравоохранения РФ от 7 сентября 2020 г. N 947н [1] ведение медицинской документации должно осуществляться в форме электронных документов согласно установленным стандартам.

В связи с этим возникает необходимость

- преобразования огромного количества уже сформированных медицинских документов, содержащих неструктурированный текст, к виду, соответствующему утвержденным шаблонам;
- реализации инструментов, позволяющих сразу создавать в МИС структурированные документы, при минимальных усилиях со стороны врача.

Эффективным инструментарием решения указанной задачи являются современные методы глубокого обучения [2], которые оказали значительное влияние на развитие искусственного интеллекта, в том числе, в области обработки естественного языка (Natural Language Processing, NLP) [3]. Особенно перспективным с точки зрения извлечения значимой информации из неструктурированного текста представляется применение больших языковых моделей (Large Language Model, LLM). Это направление демонстрирует значительный потенциал как для оптимизации различных бизнес-процессов [4], так и процессов в сфере здравоохранения, в частности [5].

Проблема исследования. Заполнение электронных медицинских документов занимает значительную часть времени приема, отвлекая врача от непосредственного общения с пациентом и тщательного осмотра. В условиях перегруженности медицинских работников это становится особенно критичным. Ситуацию часто усугубляет медленная работа программного обеспечения, превращая рутинное заполнение полей в утомительную процедуру. Сокращение времени на осмотр пациента и беседу с ним, безусловно, негативно сказывается на качестве медицинской помощи.

Для решения указанной проблемы в настоящем исследовании была поставлена цель: сократить трудоемкость формирования медицинских документов путем создания инструментария автоматического структурирования медицинских текстов с распределением данных по полям электронного документа установленного образца. Разрабатываемая система должна

«уметь» обрабатывать неструктурированный текст, полученный, в том числе, посредством распознавания голосового ввода — из единого описания осмотра, вводимого голосом. Это позволит освободить врача от рутинной работы с формами и сфокусироваться на непосредственном взаимодействии с пациентом. Кроме того, создание описанного инструментария предоставит возможность структурировать и интегрировать в действующую электронную базу данных созданные ранее неструктурированные медицинские документы, что обеспечит более полный доступ врача к медицинской истории пациентов.

Для реализации модели автоматического структурирования текста возможны два способа:

- использование предварительно обученной языковой модели с дообучением ее на собственных медицинских данных;
- использование готового LLM-сервиса.

Первый вариант требует больших вычислительных и временных ресурсов на обучение, поэтому в рамках данного исследования решено было использовать готовый LLM-сервис. При этом неизвестно, насколько успешно могут доступные LLM-сервисы справиться со специфическими медицинскими текстами и шаблонами. В связи с этим возникает задача тестирования сервисов на имеющихся медицинских данных и выбор наиболее приемлемого варианта, пригодного для практического использования в условиях медицинских учреждений.

Таким образом, в данной работе были поставлены следующие задачи.

1. Изучить рынок доступных LLM-сервисов.
2. Сформировать процедуру сравнительного анализа доступных LLM-сервисов:
   2.1) сформировать перечень критериев для сравнения;
   2.2) отобрать список LLM-сервисов для подробного изучения;
   2.3) определить набор метрик для оценки качества моделей.
3. Выполнить окончательный выбор LLM сервиса:
   3.1) сформировать тестовый набор данных;
   3.2) выполнить оценку качества LLM на тестовых данных и обосновать окончательный выбор.
4. Спроектировать архитектуру подсистемы автоматического структурирования документов.
5. Разработать механизм взаимодействия между МИС и LLM-сервисом.

Материалы и методы. В рамках данного исследования рассматривались следующие популярные российские LLM-сервисы:

- GigaChat [6];
- Яндекс GPT [7];
- T-Pro [8];
- Cotype [9].

Предварительное сравнение этих сервисов выполнялось по критериям

- языковая специализация,
- возможные области применения,
- доступность и возможность интеграции,
- размер контекста,
- наличие бесплатного тарифа API.

Для окончательной оценки качества LLM-сервисов использовались следующие метрики качества.

1. BLEU (BiLingual Evaluation Understudy) — метрика, основанная на вычислении совпадений n-грамм (последовательностей из n слов, обычно от 1 до 4) между сгенерированным результатом модели и эталоном.

   Значение BLEU вычисляется по формуле:

$$\text{BLEU} = \text{BP} \cdot \exp(\sum \omega_n \log p_n), \tag{1}$$

где BP — штраф за краткость, ω_n — весовые коэффициенты для точности n-грамм (обычно устанавливаются равными 1/N), p_n — точность для n-грамм;

точность для n-грамм вычисляется как отношение количества встречающихся последовательностей сгенерированного результата в эталонном к общему числу последовательностей сгенерированного текста:

$$p_n = \frac{\text{количество совпадающих n-грамм}}{\text{общее количество n-грамм в сгенерированном результате}} \tag{2}$$

Штраф за краткость предназначен для предотвращения завышения оценки за счет краткости сгенерированного текста:

$$\begin{cases} 1 & c > r \\ e^{1-\frac{r}{c}} & c \leq r \end{cases}, \tag{3}$$

где c — длина сгенерированного текста, r — длина эталонного текста.

2. NLI (Natural Language Inference) — метод, использующий LLM модель для оценки генерации текста. Модель принимает на вход сгенерированный и эталонный текст и определяет вероятность, насколько сгенерированный результат семантически соответствует эталонному, даже если их формулировки различаются.
3. exact_match — процент полностью совпадающих сгенерированных и эталонных значений:

$$\text{exact}_{\text{match}} = \frac{\text{количествоточныхсовпадений}}{\text{общееколичествопримеров}} \cdot 100\% \tag{4}$$

Оценка качества языковых моделей выполнялась для медицинских документов наиболее часто используемого типа — «Протокол консультации (CDA) Редакция 5» [10], выгруженных из медицинской информационной системы (МИС) «1С:Здравоохранение 72» [11]. Для выгрузки данных в системе была реализована обработка, на форму которой вынесены следующие параметры для отбора документов:

- период дат (для отбора более актуальных документов);
- ШМД (для указания конкретного вида документов);
- авторы документов (для указания врачей, которые заполняют документы более корректно и правильно).

Отобранные медицинские документы были выгружены в формате XML, для их последующей обработки написан скрипт на языке Python, реализующий парсинг XML-файлов с итеративным обходом их структуры для извлечения полей и соответствующих им значений.

Для взаимодействия с сервисом GigaChat были использованы методы его API (см. табл. 1); вызов методов осуществлялся путем формирования и отправки HTTP-запросов. Ключ

авторизации и токен доступа передавались в заголовках запроса, текстовое сообщение — в теле запроса в формате JSON. Ответы от API возвращаются также в JSON, что обеспечивает легко обрабатываемый формат данных.

Таблица 1

Методы GigaChat API

Наименование метода	*Описание*	*Входные данные*	*Возвращаемое значение*
oauth	Возвращает токен доступа	Ключ авторизации	Токен доступа
chat/completions	Возвращает ответ модели, сгенерированный на основе переданного сообщения	Токен доступа; текст сообщения	Ответ
balance	Возвращает доступный остаток токенов	Токен доступа	Остаток токенов

Результаты. Результаты предварительного сравнения четырех выбранных LLM-сервисов представлены в табл. 2.

Таблица 2

Сравнительная таблица LLM сервисов

Критерий сравнения	*GigaChat*	*Яндекс GPT*	*T-Pro*	*Cotype*
Языковая специализация	Русский язык	Русский язык	Русский язык	Русский язык
Область применения	Широкий круг задач	Широкий круг задач	Бизнес-задачи	Бизнес-задач
Интеграция и доступность	Доступен через платформу Сбербанка и API-интерфейсы, удобен для интеграции в различные системы и приложения	Интегрирован в сервисы Яндекса и доступен для сторонних разработчиков через API платформы Yandex.Cloud	Предназначен для внутренних компаний-клиентов и крупных корпораций, доступна интеграция в рабочие процессы предприятий	Предназначен для внутренних компаний-клиентов и крупных корпораций, доступна интеграция в рабочие процессы предприятий
Размер контекста	32768 токенов	2000 токенов		
Наличие бесплатного тарифа API	+ (1 000 000 токенов на год)	-	-	-

Поскольку T-Pro и Cotype более ориентированы на решение бизнес-задач и не предоставляют бесплатного доступа для тестирования, в дальнейшем рассматривались только GigaChat и Яндекс GPT.

Для тестирования моделей был сформирован промт, позволяющий передать исходный текст, список необходимых полей, некоторые уточняющие инструкции и получить в качестве ответа значения полей в формате JSON:

«*Извлеки из предоставленного текста информацию по следующим полям: <перечисление полей>.*

Представь результат в формате JSON

[

<поле>: <значение>

...

<поле>: <значение>

]

Учти:

- *если для поля информация не найдена, то оставь значение пустым;*
- *значения полей [Основной диагноз, Предварительный диагноз] должны быть в формате МКБ-10.*

Исходный текст для обработки: <текст>».

Для получения достоверных результатов сравнения качества работы сервисов тестирование проводилось на одном и том же наборе документов. Поскольку Яндекс GPT не имеет бесплатного тарифа API, тестирование этого сервиса пришлось выполнять вручную, что привело к вынужденному сокращению количества обрабатываемых документов до 50.

Значения метрик качества, полученные по результатам тестирования, представлены в табл. 3.

Таблица 3

Результаты тестирования

	GigaChat			Яндекс GPT		
	exact_match	bleu	nli	exact_match	bleu	nli
Анамнез жизни	0%	0,065	0,711	0%	0,192	0,734
Анамнез заболевания	0%	0,038	0,718	0%	0,204	0,817
Вес	90%	0,937	0,993	100%	1,000	0,994
Детализация основного диагноза	0%	0,094	0,408	0%	0,105	0,493
Жалобы	0%	0,194	0,553	0%	0,244	0,637
Заключение	0%	0,000	0,151	0%	0,000	0,168
Объективный статус	0%	0,114	0,803	0%	0,152	0,802
Основной диагноз	20%	0,305	0,908	90%	0,900	0,898
Рекомендации	0%	0,137	0,420	0%	0,176	0,382
Рост	90%	0,937	0,982	100%	1,000	0,996
Сопутствующие диагнозы	0%	0,000	0,321	0%	0,000	0,199

В целом значения метрик оказались сравнимыми у обоих сервисов. Однако в процессе тестирования выяснилось, что размера контекста Яндекс GPT не хватает для полноценной обработки имеющихся реальных документов, вследствие чего сгенерированные ответы

иногда «обрезаются». Кроме этого, существенным недостатком этого сервиса является отсутствие бесплатного тарифа API. Поэтому для дальнейшей работы в качестве основного сервиса был выбран GigaChat.

Общая архитектура подсистемы автоматического структурирования информации из медицинских документов представлена на рис. 1.

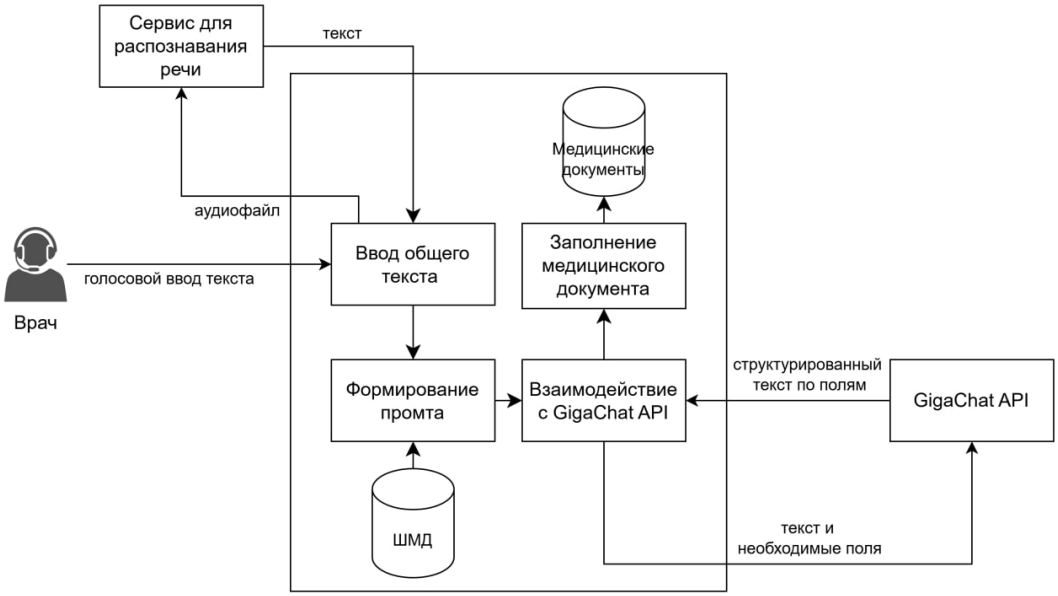

Рис. 1. Общая архитектура подсистемы

В конфигурации МИС реализовано расширение, включающее новый объект конфигурации — обработку, интерфейс которой содержит поле для ввода неструктурированного медицинского текста. В этой обработке реализована логика формирования необходимого промта для GigaChat на основе введенного текста и перечня полей электронного медицинского документа и методы для осуществления взаимодействия с GigaChat API.

На рис. 2 представлен пример неструктурированного текста о результатах осмотра, введенного во входное поле обработки, а на рис. 3 и 4 — результаты структурирования — распределенные по заданным полям фрагменты текста.

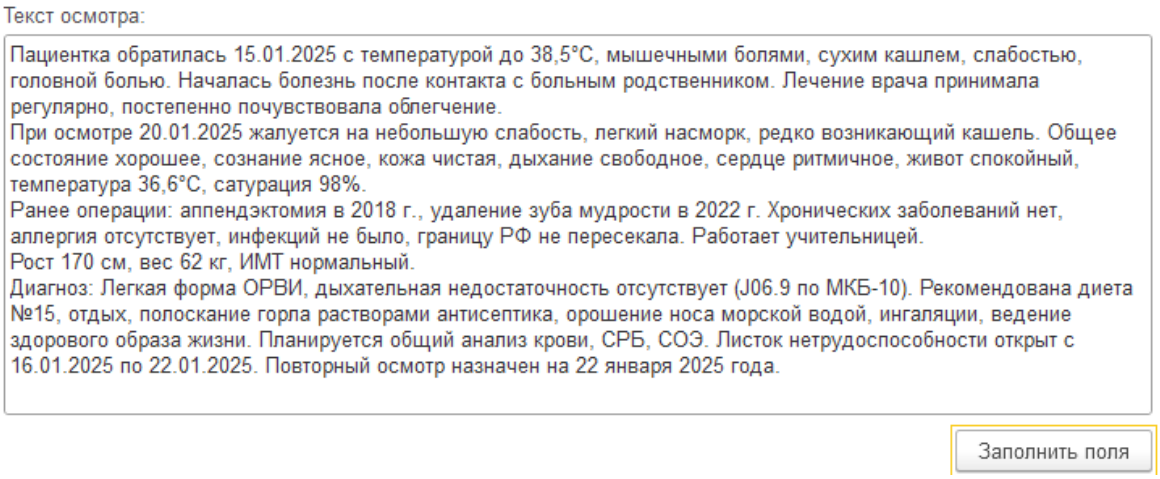

Рис. 2. Поле для ввода общего текста

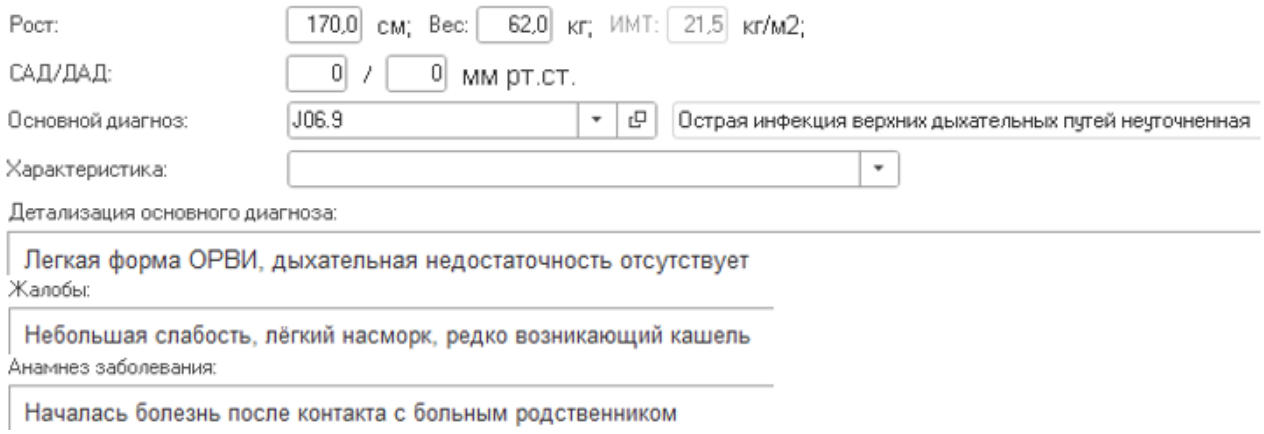

Рис. 3. Заполненные поля

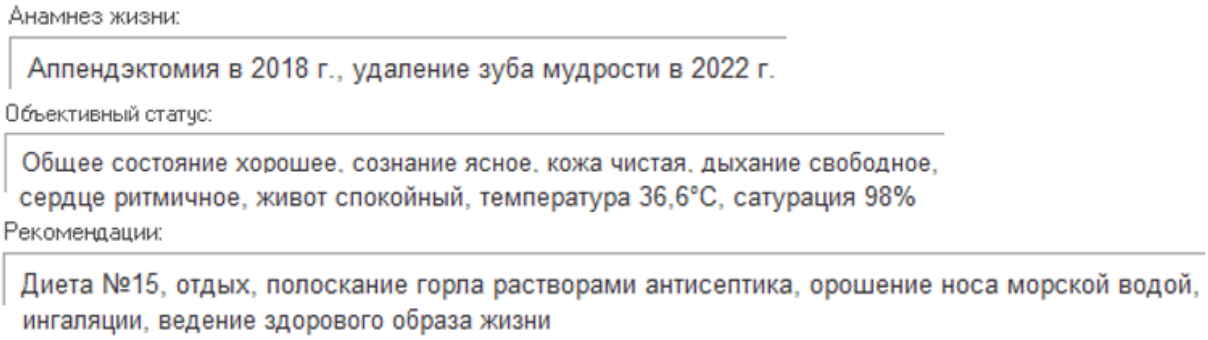

Рис. 4. Заполненные поля

Заключение. Разработана подсистема автоматического структурирования информации из медицинских документов на основе LLM-сервиса GigaChat. Сравнительный анализ показал сопоставимую эффективность GigaChat и Яндекс GPT, однако практические ограничения последнего привели к выбору GigaChat в качестве основы для решения. Предполагается, что разработанная подсистема позволит эффективно извлекать информацию из свободного текста и заполнять поля электронных медицинских форм, сокращая время работы медицинского персонала и повышая эффективность их труда.

СПИСОК ЛИТЕРАТУРЫ

1. Официальное опубликование правовых актов Приказ Министерства здравоохранения Российской Федерации от 07.09.2020 № 947н "Об утверждении Порядка организации системы документооборота в сфере охраны здоровья в части ведения медицинской документации в форме электронных документов" (Зарегистрирован 12.01.2021 № 62054) — URL: http://publication.pravo.gov.ru/Document/View/0001202101120007 (дата обращения 05.05.2025).
2. Шолле Франсуа Глубокое обучение на Python. — Санкт-Петербург : Питер, 2018. — 400 с.: ил. — (Серия «Библиотека программиста»).
3. Сметана, В. В. Глава 13. Глубокое обучение и нейронные сети: новая эра искусственного интеллекта / В. В. Сметана // Теория и практика конфликта. Очерки современной конфликтологии. — Санкт-Петербург : Частное научно-образовательное учреждение дополнительного профессионального образования Гуманитарный национальный исследовательский институт «НАЦРАЗВИТИЕ», 2024. — С. 76-78. — EDN KHEYKK.

4. Молчанов, И. А. Оптимизация бизнес-процессов за счет использования LLM-моделей / И. А. Молчанов, В. М. Джуха // Научный вектор : Сборник научных трудов. — Ростов-на-Дону : Ростовский государственный экономический университет "РИНХ", 2024. — С. 280-284. — EDN QGUUVD.
5. Назаров, Д. М. Сущность применения LLM моделей в сфере здравоохранения / Д. М. Назаров, Ф. И. Бадаев // Искусственный интеллект в решении актуальных социальных и экономических проблем XXI века : Сборник статей по материалам Девятой всероссийской научно-практической конференции с международным участием, Пермь, 17–18 октября 2024 года. — Пермь: Пермский государственный национальный исследовательский университет, 2024. — С. 257-261. — EDN MCYXKS.
6. GigaChat — русскоязычная нейросеть от Сбера : [сайт]. — URL: https://giga.chat/ (дата обращения: 05.05.2025).
7. YandexGPT 5 — новое поколение нейросетей Яндекса : [сайт]. — URL: https://ya.ru/ai/gpt (дата обращения: 05.05.2025).
8. T-Bank AI | AI-powered FinTech в Т-Банке : [сайт]. — URL: https://ai.tbank.ru/ (дата обращения: 05.05.2025).
9. Генеративный ИИ — Cotype : [сайт]. — URL: https://mts.ai/ru/product/generative-ai-solutions/ (дата обращения: 05.05.2025).
10. Портал оперативного взаимодействия участников ЕГИСЗ: [сайт]. — URL: https://portal.egisz.rosminzdrav.ru/materials (дата обращения 05.05.2025).
11. 1С-Медицина-Регион. Управление ресурсами медицинских организаций: [сайт]. — URL: https://1cmr.ru/ (дата обращения 05.05.2025).