Система автоматического заполнения анамнеза

Т. М. Татарникова

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

tm-tatarn@yandex.ru

В. В. Цехановский

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

vvcehanovsky@mail.ru

Б. Я. Советов

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

bysovetov@mail.ru

Д. Р. Миляев

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

milyaev.dmitry00@mail.ru

Аннотация. **Обсуждается задача автоматического заполнения анамнеза на основе диалога врача и пациента. Приведены предложения по технической реализации медицинской информационной системы с микросервисной архитектурой, включающей сервисы по распознаванию речи, формированию анамнеза, регистрации пациента. Показано, что задача актуальна в условиях ограниченности ресурсов и необходимости работы с чувствительными медицинскими данными без выхода в интернет. Приведены обоснования для выбора технологий и инструментов распознавания речи и моделей обработки естественного языка для интеграции в медицинскую систему.**

Ключевые слова: автоматизация заполнения анамнеза; диалог врача и пациента; распознавание речи; транскрибация; медицинская информационная система; микросервисная архитектура

I. Введение

Развитие технологий распознавания речи открыло широкие возможности для автоматизации рутинных задач в медицине, таких как заполнение медицинской документации [1]. На практике это позволяет сократить время приёма пациента и снизить нагрузку на медицинский персонал. Тем не менее, большинство существующих решений имеют узкую специализацию и не охватывают весь необходимый функционал, особенно когда речь идет о полноценной автоматической генерации отчётной документации на основе диалога между врачом и пациентом

В работе выдвигается гипотеза о возможности технической реализации системы распознавания и обработки речи для автоматической записи диалогов между врачом и пациентом, а также последующим её анализом и составлением отчётного документа.

II. Анализ существующих решений

A. Коммерческие решения

На рынке представлено несколько коммерческих решений медицинских информационных систем (МИС), предназначенных для автоматизации медицинской документации с использованием технологий распознавания речи [2].

Одним из наиболее известных является Nuance Dragon Medical One – облачное решение, позволяющее преобразовывать устную речь в текст с высокой точностью, поддерживает медицинские словари и адаптировано под терминологию клинической практики. Благодаря этим функциям система широко используется в странах с англоязычной медицинской документацией.

Voice2Med – платформа, предлагающая инструменты для голосового ввода медицинских данных. Она также ориентирована на диктовку и ускорение оформления записей, однако акцент делается исключительно на монолог врача, без поддержки обработки диалогов с пациентом.

Подобные решения помогают ведению документации, но их функциональность ограничивается преобразованием речи в текст и не включает полноценную структуризацию информации или интеграцию с системами анализа естественного языка. Более того, отсутствие поддержки русского языка в большинстве таких систем делает их малоприменимыми в отечественной практике.

B. Технологии и инструменты

Существующие в настоящее время решения МИС в основном представляют собой различное проприетарное программное обеспечение для распознавания голоса и обработки естественного языка, среди которых [3]:

Yandex SpeechKit – облачный сервис от Яндекса, предоставляющий технологии распознавания и синтеза речи. Сервис позволяет преобразовывать устную речь в текст и наоборот, что удобно для создания голосовых интерфейсов, виртуальных ассистентов, чат-ботов и систем автоматического документооборота. Сервис поддерживает русский язык с высокой точностью и может работать как с заранее записанными аудиофайлами, так и в режиме реального времени.

AWS Transcribe Medical – сервис для преобразования речи в текст, ориентированный на медицинскую терминологию. Сервис в основном рассчитан на диктовку со стороны врача и не поддерживает автоматическую обработку диалогов.

Microsoft Dragon Copilot – продукт, сочетающий распознавание речи с искусственным интеллектом, но пока остаётся ограниченным в возможностях анализа взаимодействий между врачом и пациентом, сосредотачиваясь на ассистировании в документации по команде.

LiveMedical – система, направленная на точность распознавания медицинских терминов. Система не включает в себя модули глубокого анализа текста или автоматического формирования медицинских заключений.

Google Speech-to-Text API – универсальное решение для распознавания речи, включая медицинские сценарии. Приложение требует значительной доработки и интеграции дополнительных компонентов для полноценного использования в сфере здравоохранения.

III. Описание архитектуры программного обеспечения

Реализация МИС соответствует концепции микросервисной архитектуры, что обусловлено рядом причин:

- модульность: каждый компонент системы реализует строго определённую функцию и может разрабатываться, тестироваться и масштабироваться независимо;
- гибкость в развитии: позволяет заменить или доработать отдельный сервис без вмешательства в другие сервисы системы;
- изоляция отказов: сбой одного сервиса не влияет на работу остальных модулей.
- масштабируемость: возможность масштабировать только те сервисы, которые испытывают нагрузку;
- безопасность: каждый сервис может быть изолирован и настроен на работу в закрытом контуре локальной сети.

Структура МИС представлена на рис. 1. Список сервисов и их описание представлены в табл. I.

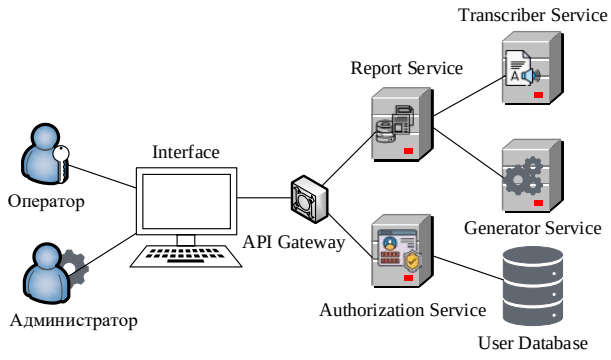

Рис. 1. Архитектура системы

ТАБЛИЦА I. Описание сервисов МИС

Сервис	Описание
Interface (Desktop/UI)	Графический интерфейс для взаимодействия пользователя: оператора или администратора с системой
API Gateway	Единая точка входа для всех клиентских запросов
Authorization Service	Микросервис, отвечающий за хранение и проверку логинов/паролей, а также ролей пользователей (оператор/администратор)
Transcriber Service	Микросервис для распознавания речи (Speech-to-Text) из загруженных аудиофайлов
Generator Service	Микросервис, использующий большую языковую модель (LLM) для структурирования текста и извлечения медицинских сущностей
Report Service	Микросервис для формирования финального документа по заданному шаблону

Общение сервисов происходит по REST API (Representational State Transfer Application Programming Interface) – архитектурный стиль взаимодействия между клиентом и сервером через HTTP. Каждый сервис отвечает за строго ограниченную бизнес-логику, а для обмена информацией используются унифицированные HTTP-запросы и структурированные данные в формате JSON или бинарные данные, например, аудиофайлы и DOCX-документы.

IV. Этапы распознования речи

Процесс распознавания речи состоит из нескольких основных этапов:

Предобработка сигнала: речевой сигнал очищается от шума и разбивается на последовательность фреймов. Каждый фрейм представляет собой небольшой отрезок сигнала, на основе которого вычисляются акустические признаки.

Акустическая модель: связывает речевые признаки с фонемами – наименьшими звуковыми единицами языка. Формально, цель модели – найти максимальное правдоподобие последовательности наблюдений X, определяемое как:

$$P(X \mid W) = \prod_{i=1}^{T} P(x_i \mid s_i),$$

где W – последовательность слов; x_i – наблюдения (фреймы); s_i – состояния (например, фонемы или их комбинации).

Лингвистическая модель: оценивает вероятность последовательности слов на основе статистики встречаемости и грамматических закономерностей. Например, для предсказания следующего слова в последовательности могут использоваться N-граммные модели, которые оценивают вероятность появления слова с учетом предшествующих

$$P(w_n \mid w_{n-1}, w_{n-2}, \ldots, w_{n-N+1}).$$

Более современные методы, такие как трансформеры и рекуррентные нейронные сети, также могут моделировать длинные зависимости между словами.

Декодирование: на основе акустической и лингвистической моделей декодер выбирает последовательность слов W, которая наиболее вероятно соответствует входному сигналу

$$W = \arg\max_{W} P(W \mid X) = \arg\max_{W} P(X \mid W) P(W).$$

V. Анализ технологий

Распознавание речи (транскрибация) выполняется с применением различных библиотек и платформ. В работе проведен анализ open source моделей, разработанные крупными компаниями, с активной поддержкой и подробной документацией. Поэтому были выбраны четыре семейства моделей транскрибации (табл. II).

ТАБЛИЦА II. Модели транскрибации

Семейство моделей	Конфигурация	№ модели	Разработчик	Лицензия
Whisper	Whisper Tiny	1	OpenAI	MIT License
	Whisper Base	2		
	Whisper Small	3		
	Whisper Medium	4		
	Whisper Large	5		
GigaAM	GigaAM-CTC	6	Сбер	NonCommercial Public License
	GigaAM-RNNT	7		
Vosk	Vosk Small	8	AlphaCep	Apache License 2.0
	Vosk Big	9		
Nemo	Nemo STT Ru Conformer-CTC Large	10	Nvidia	NGC License для Conformer-Transducer Large конфигурации CC BY 4.0 для остальных конфигураций
	Nemo STT Ru Conformer-Transducer Large	11		
	Nemo STT Multilingual FastConformer Hybrid Transducer-CTC Large P&C	12		

При выборе конкретного семейства и конфигурации учитывались следующие факторы: аппаратные ограничения; ресурсы и вычислительная эффективность; открытость лицензии; точность распознавания; возможность дообучения.

Таким образом, из дальнейшего рассмотрения были исключены модели 6 и 7 из-за невозможности дообучения, модель 11 по причине того, что NGC License запрещает использование программного обеспечения в продуктах или услугах, предназначенных для критически важных приложений.

Все выбранные модели протестированы на аудиодатасетах, включающие разнообразные примеры речи, такие как «Ru Librispeech», «Команды Яндекса», «Медицина» и др. [4].

Выбор модели транскрибации выполнен на основе сравнительного анализа по следующим метрикам.

Word Error Rate (*WER*) – оценка точности распознавания речи, показывает процент неправильно распознанных слов в сравнении с эталонным текстом

$$WER = \frac{S + D + I}{N},$$

где S – количество замен слов; D – количество удалений слов, I – количество вставок слов; N – общее количество слов в эталонном тексте.

Edit Distance (*ED*) – минимальное количество операций (вставка, удаление, замена), необходимых для преобразования одного текста в другой

$$ED = \frac{\sum_{i=1}^{n} d(R_i, H_i)}{n},$$

где d – расстояние редактирования между эталоном R и гипотезой H; n – общее количество пар текстов.

Real-Time Factor (*RTF*) – отношение времени обработки аудиофайла к его общей длительности

$$RTF = \frac{T_{\text{Process}}}{T_{\text{Audio}}},$$

где T_{Process} – время обработки аудиофайла; T_{Audio} – длительность аудиофайла.

Суммарное время работы, t_{Σ} – общее время, затраченное моделью на обработку всех аудиофайлов. Оценивает общую производительность модели на выбранном датасете

$$t_{\Sigma} = \sum_{i=1}^{k} T_{\text{Process}\_i},$$

где T_{Process_i} – время обработки i-го аудиофайла; k – количество аудиофайлов.

Среднее время обработки на файл, $\bar{t}$ – среднее время обработки одного аудиофайла моделью. Оценивает эффективность модели при обработке единичного файла

$$\bar{t} = \frac{t_{\Sigma}}{k}.$$

Распределение ошибок по длине слов – распределение частоты ошибки распознавания в зависимости от длины слов в эталонном тексте. Позволяет понять, на каких словах (коротких или длинных) модели чаще допускают ошибки.

Результаты тестирования приведены на рис. 2–7.

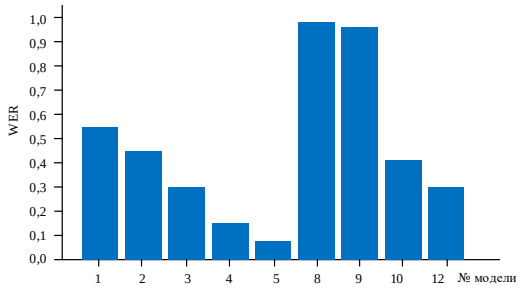

Рис. 2. Оценка точности распознавания речи WER

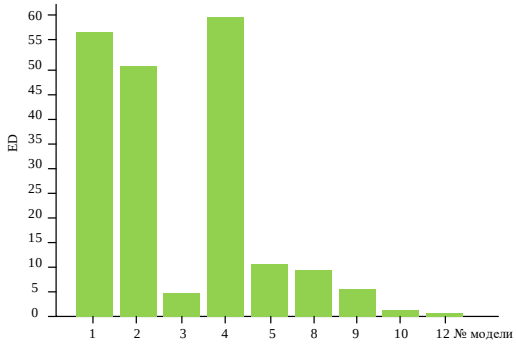

Рис. 3. Оценка среднего расстояния редактирования *ED*

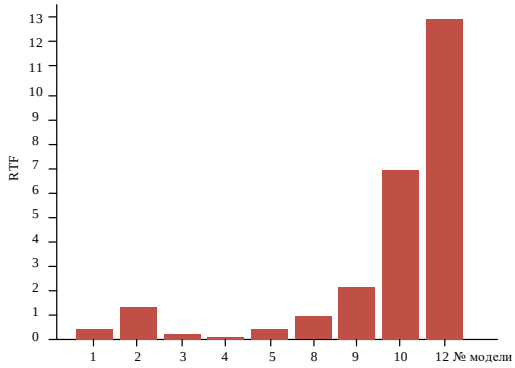

Рис. 4. Оценка RTF

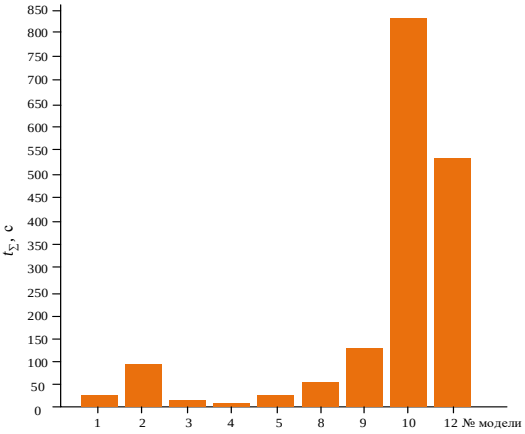

Рис. 5. Оценка суммарного времени работы

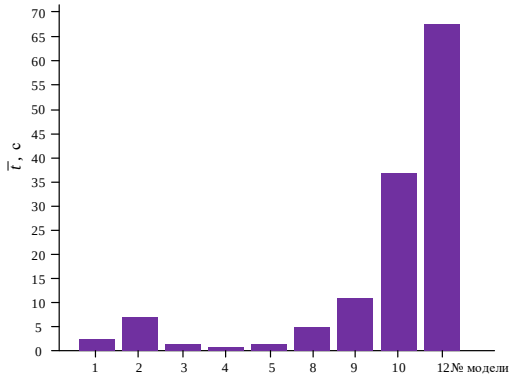

Рис. 6. Оценка среднего времени обработки аудиофайла

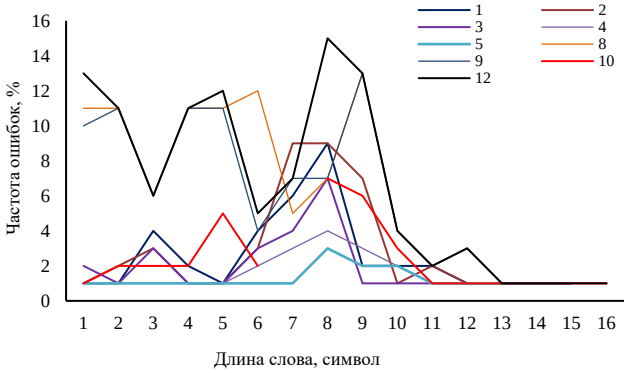

Рис. 7. Распределение ошибок по длине слов

После распознавания речи требуется применение моделей обработки естественного языка (Natural Language Processing, NLP), способных интерпретировать, классифицировать и структурировать диалоги между врачом и пациентом. В задаче приоритет также отдан локальным open source языковым моделям, не требующим подключения к интернету и поддерживающим автономную работу в изолированной среде медицинского учреждения.

Рассмотрены следующие семейства современных больших языковых моделей, представленных в открытом доступе и совместимых с фреймворками для локального запуска: Mistral 7B, GPT-J 6B, Phi-2, Falcon, и RWKV, GPT-4 / GPT-3.5. Проведенный анализ и тестирование этих моделей позволили выбрать Mistral 7B, поскольку модель способна работать локально и имеет обширную подробную документацию для установки и настройки.

Предусловием работы программы является наличие размеченных шаблонов отчётов в хранилище. Для этого администратор должен получить файл отчёта и разметить его особым образом.

VI. Заключение

В перспективе проект может быть существенно расширен и интегрирован в более широкие медицинские информационные системы. Возможные направления развития включают:

- интеграцию с существующими МИС/ЭМК — возможностью напрямую экспортировать отчёты в базы данных медицинских учреждений;
- добавление аналитики и статистики — формирование отчётов по динамике приёмов, частым жалобам, автоматизированное заполнение статистических форм;
- интеграцию с мобильными устройствами — возможность диктовать текст и получать отчёты с мобильных устройств, планшетов, интеграция с цифровыми диктофонами;
- постпроцессинг и дообучение — улучшение модели с помощью дообучения и замены ошибочных терминов.
- расширение перечня шаблонов и специалистов — добавление новых размеченных отчётов и возможность использовать ПО широким кругом специалистов.

Список литературы

[1] Huha J., Parka S., Leeb J. E., Ye J. C. Improving Medical Speech-to-Text Accuracy with Vision-Language Pre-training Model // IEEE J Biomed Health Inform. 2024. P. 1692-1703.

[2] Santhosh D., Nataraj K. S., Nitya T. Self-Training and Error Correction using Large Language Models for Medical Speech Recognition // 2024 IEEE Conference on Engineering Informatics (ICEI). 2024. P. 1-6. DOI:10.1109/ICEI64305.2024.10912300/

[3] Valuitseva I., Filatov I. Design of ASR Software for Recognition of the Russian Language Variants // Polylinguality and Transcultural Practices. 2021. Vol.18. P. 245-254. DOI: 10.22363/2618-897X-2021-18-3-245-254.

[4] Открытые модели для распознавания русской речи 2024. URL: https://alphacephei.com/nsh/2024/04/14/russian-models.html (дата обращения 23.02.2026)