ISSN 2782-2001
Системы анализа
и обработки данных
том 103, № 3, 2026, с. 75–96

http://journals.nstu.ru/vestnik
Analysis and data
processing systems
Vol. 103, No. 3, 2026, pp. 75–96

ИНФОРМАЦИОННЫЕ
ТЕХНОЛОГИИ
И ТЕЛЕКОММУНИКАЦИИ

INFORMATION
TECHNOLOGIES
AND TELECOMMUNICATIONS

УДК 519.85

DOI: 10.17212/2782-2001-2026-3-75-96

Обработка текстовых данных в медицинских диалоговых системах на примере виртуального ассистента эндокринолога*

М.С. МАМИКЬЯН[a], Я.А. БЕКЕНЕВА[b]

РФ, 197022, г. Санкт-Петербург, ул. Профессора Попова, 5, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[a] *mmamikyan2001@gmail.com* [b] *yabekeneva@etu.ru*

Доля больных с патологиями эндокринной системы постоянно увеличивается во всех экономически развитых странах мира. В России за последние 6 лет распространенность заболеваний эндокринной системы выросла на 26 %, а показатели первичной заболеваемости – на 36 %. Одной из главных проблем, усугубляющих эпидемиологическую ситуацию, является дефицит местных терапевтов и неравномерное распределение медицинских специалистов между сельскими и городскими регионами. Жители малонаселенных пунктов вынуждены ездить в соседние города для получения необходимой медицинской помощи, что влечет дополнительные временные и финансовые затраты.

Также стоит отметить, что определение необходимых анализов для сдачи может занять немало времени, а спешка при приеме может привести к сдаче избыточных анализов. Для оптимизации этого процесса важно заранее определить, какие именно анализы нужны пациенту, что требует постановки предварительного диагноза. В этом может помочь система распознавания заболеваний на основе жалоб пациента, которая сможет автоматизировать постановку предварительного диагноза и избежать ненужных исследований. Основу представленного программного решения составляет комбинация использования распознавания именованных сущностей для извлечения симптомов из жалоб пациентов в свободной форме и методов машинного обучения для определения диагноза. Для обучения использован набор данных RuDRec в совокупности с синтетическими данными. Описан алгоритм диалога с пользователем для уточнения симптомов и формирования перечня необходимых анализов.

Такая система поможет сократить случаи самолечения, когда пациенты, не имея точных рекомендаций, следуют советам знакомых. Таким образом, формирование направлений на анализы на основе качественной первичной консультации пациентов, особенно в малонаселенных пунктах, является актуальной задачей. Практическая значимость работы заключается в оптимизации нагрузки на медицинские учреждения, в снижении рисков самолечения и повышении диагностики в удаленных регионах.

* *Статья получена 06 июня 2025 г.*

Ключевые слова: виртуальный ассистент эндокринолога, методы обработки медицинских данных, большие языковые модели, чат-бот, диалоговый интерфейс, диагностика, эндокринологическое заболевание, телемедицина, именованные сущности, распознавание текстов, метрики точности, интеллектуальная система машинного обучения

ВВЕДЕНИЕ

Постановка эндокринного диагноза представляет собой комплексную проблему. В монографиях по эндокринным заболеваниям, как правило, методы исследования эндокринной системы описаны схематично [1]. Диагностика пациента включает в себя следующее [2].

1. Анамнез. Жалобы у больных с эндокринной патологией чрезвычайно многообразны, поэтому опрос пациента должен включать в себя несколько этапов, которые позволят врачу собрать информацию о симптомах и их продолжительности. Врач тщательно выслушивает все жалобы пациента и задает уточняющие вопросы, чтобы сузить круг возможных заболеваний. Затем уточняется история жизни больного, особое внимание уделяется опросу о родственниках, так как ко многим эндокринопатиям имеется наследственная предрасположенность.

2. Осмотр. При общем осмотре пациента врач обращает внимание на несколько ключевых аспектов: внешний вид, состояние кожи и волос, особенности лица.

3. Пальпация – физический метод медицинской диагностики, проводимый путем ощупывания тела пациента. При пальпации щитовидной железы врач ориентируется на ее размер, консистенцию и наличие узлов.

После сбора анамнеза, исходя из жалоб и состояния пациента, врач может поставить предполагаемый диагноз и назначить определенные анализы для постановки точного диагноза. Перечень анализов подбирается индивидуально врачом-эндокринологом в зависимости от симптомов, предварительного диагноза и состояния пациента. Автоматизация процесса сбора анамнеза и подбора необходимых анализов может позволить сократить количество приемов пациента у медицинского специалиста. Такое решение может стать особенно актуальным для отдаленных населенных пунктов, где могут отсутствовать профильные специалисты. В некоторых регионах России время ожидания приема профильного специалиста может достигать нескольких месяцев.

Технологии телемедицины сегодня являются важным инструментом для обеспечения равного доступа к медицинской помощи по всему миру. Их активное внедрение уже стало стандартом во многих странах, включая Россию. Электронная запись к врачу, система электронной очереди и другие цифровые сервисы помогают повысить качество обслуживания пациентов, оптимизировать работу медицинского персонала и снизить затраты на оказание медицинских услуг. Современные телемедицинские технологии охватывают широкий спектр задач – от диагностики и лечения заболеваний до дистанционного биомониторинга [3]. Перспективы развития телемедицины связаны с дальнейшим внедрением цифровых технологий. Уже сейчас телемедицинские платформы активно используются для удаленных консультаций с узкими специалистами, для мониторинга хронических заболеваний и контроля жизненно важных показателей, таких как уровень сахара в крови, артериальное давление, сердечный ритм. Несмотря на активное развитие, имеются факторы, ограничивающие

широкомасштабное внедрение телемедицины. Тем не менее мировая практика показывает, что в ближайшие годы телемедицина станет неотъемлемой частью системы здравоохранения, обеспечивая доступность и качество медицинских услуг независимо от географического положения пациента [4].

В первом разделе статьи рассматриваются и анализируются существующие постановки задачи обработки текстов и создания диалоговых систем. Во втором разделе представлен подход к обработке жалоб, вводимых пациентом в текстовой форме. В третьем разделе представлены результаты экспериментальных исследований предлагаемой диалоговой системы.

1. ПОСТАНОВКА ЗАДАЧИ

Задача автоматизации процесса первичной диагностики пациента может быть сведена к задаче автоматизации сбора анамнеза. Как правило, на приеме у врача-эндокринолога пациент перечисляет жалобы в свободной форме, а врач задает уточняющие вопросы. Автоматизация этого процесса может быть реализована в виде диалоговой системы. Чат-боты, например, становятся неотъемлемой частью цифровой медицины, помогают автоматизировать взаимодействие с пациентами и снижать нагрузку на медицинские учреждения. Они представляют собой программных ассистентов, способных анализировать жалобы, задавать уточняющие вопросы и предоставлять рекомендации. Первые подобные системы появились еще в 1960-х годах, однако только в последние годы они активно внедряются в сферу здравоохранения.

1.1. ЧАТ-БОТЫ КАК СРЕДСТВО ОРГАНИЗАЦИИ ДИАЛОГА С ПАЦИЕНТОМ

Современные медицинские диалоговые системы, в частности чат-боты, выполняют различные функции – от поиска ближайших медицинских учреждений и записи на прием к врачу до сбора первичной информации о симптомах пациента. Такие сервисы доступны бесплатно и могут оценивать необходимость визита к врачу. Они помогают экономить время, могут напоминать пациентам о приеме лекарств, предоставлять рекомендации и минимизировать число излишних обращений в медицинские учреждения. Чат-бот может проконсультировать пациента в экстренной ситуации, предоставив рекомендации по оказанию первой помощи до приезда специалистов [5, 6].

Babylon [7] – медицинский чат-бот, предлагающий консультацию как на основе искусственного интеллекта, так и непосредственно врачом. Эта система включает в себя технологии обработки естественного языка и алгоритмы машинного обучения. Пользователь в свободной форме вводит симптомы, затем чат-бот на основе введенных жалоб задает уточняющие вопросы. После завершения цикла опроса пациента система сравнивает полученные данные с медицинскими базами знаний и выдает вероятные диагнозы.

Ada [8] – популярный чат-бот, ориентированный на самодиагностику. Принцип работы пользователя с системой аналогичен *Babylon*. Пользователь в свободной форме вводит жалобу. Далее система предлагает список подходящих жалоб с их подробным описанием, чтобы пользователь мог выбрать наиболее соответствующую проблему. После ввода пациентом всех жалоб

система задает ряд встречных вопросов с вариантами ответов. После завершения цикла опроса пациента система выводит описание предположительного заболевания. В описание входит информация о том, сколько по статистике человек сталкивались с этой проблемой, также подробно описываются риски, симптомы, требуемый уход, профилактика и прогнозы. Стоит отметить, что *Ada* имеет возможность интегрирования в медицинские системы, предоставляя обширный функционал. Например, с его помощью можно эффективно направлять пациентов к нужным врачам, тем самым снижая нагрузку на медицинское учреждение.

Buoy Health [9] – система, обученная на основе большого массива клинических данных. Проблема, вокруг которой была сформирована эта система, заключалась в том, что люди неправильно интерпретируют свои симптомы, основываясь на информации, которую они нашли в интернете. Принцип работы аналогичен *Ada*. Существенным отличием является разная последовательность ввода жалоб. В *Ada* пользователь в первую очередь заносит все жалобы, которые его беспокоят, и только затем система задает встречные вопросы. В *Buoy Health* пользователь может ввести следующий симптом только после ответа на ряд вопросов о первом симптоме. Также стоит отметить, что в рамках каждого нового диалога *Buoy Health* задает ряд вопросов, не связанных с жалобой пациента, чтобы исключить экстренные ситуации, например приступ инсульта. В процессе опроса ассистент уточняет геолокацию, чтобы в дальнейшем порекомендовать ближайшее медицинское учреждение. После завершения цикла опроса пациента система выводит названия предположительных диагнозов, сопровождая их статьями, с которыми пользователь может ознакомиться.

Florence [10] – надстройка мессенджера, которая работает как ассистент пациента. В отличие от рассмотренных ранее систем, *Florence* не предоставляет диагностику заболеваний. *Florence* помогает пациентам соблюдать режим приема лекарств, напоминая им о необходимости принимать медикаменты, а также предоставляет информацию о назначенных препаратах; мотивирует пользователей придерживаться назначенного режима, что снижает риск ухудшения здоровья из-за неправильного или нерегулярного приема лекарств. Кроме того, *Florence* может отслеживать состояние здоровья пользователя и помогает достигать поставленных целей. Например, с помощью этого чатбота можно контролировать параметры веса. *Florence* также предоставляет информацию о различных заболеваниях и может помочь в поисках врача или аптеки.

Symptomate [11] – система быстрой диагностики симптомов. Работает аналогично *Ada* и *Buoy Health*, но имеет меньший дополнительный функционал, в открытом доступе. В отличие от рассмотренных выше систем, *Symptomate* предоставляет удобный функционал выбора зоны, беспокоящей пациента. На основе выбранной части тела система выдает список возможных жалоб, относящихся к ней. После завершения диалога с пациентом *Symptomate* выводит список возможных заболеваний. Каждому заболеванию соответствует панель инструментов, отражающая вероятность наличия этого заболевания у пациента.

Sensely [12] – виртуальный медицинский ассистент, использующий ИИ для взаимодействия с пациентами. На основе диалога с пациентом система

помогает выявить, требуется ли немедленное вмешательство врача. Интегрируясь в медицинские учреждения, *Sensely* предоставляет дополнительный функционал: помощь в записи пациентов на прием к врачу, возможность видеоконсультаций с реальными врачами, трекер состояния здоровья.

Все рассмотренные системы представляют собой инновационные инструменты, использующие искусственный интеллект для упрощения получения первичной медицинской диагностики, мониторинга здоровья и в целом повышения доступности медицинской помощи. Каждая из этих систем имеет свои преимущества и недостатки. Среди преимуществ можно выделить легкодоступность для жителей тех стран, где эти системы распространены. Пользователь имеет круглосуточный доступ к медицинской помощи, при этом снижается нагрузка на медицинские учреждения, так как система помогает эффективно распределить потоки пациентов, направляя их к нужным специалистам. Использование такой системы способствует уменьшению случаев ошибочной самодиагностики путем выявления диагноза через интернет. Однако рассмотренные системы в настоящее время недоступны для использования в России. Кроме того, они характеризуются отсутствием локализации языка, что значительно ограничивает их использование среди пациентов, не владеющих английским языком; имеют ограниченный функционал, многие платформы не предоставляют в открытый доступ полный функционал. Среди существующих проблем выделяют ограничения, связанные с данными, на которых обучаются модели [13], этические аспекты использования подобных систем, возможное неправильное толкование полученной информации пациентами. Отмечается также, что популярные чат-боты дают общие рекомендации без учета состояния пользователя и особенностей его показателей здоровья [14].

Чат-боты не могут самостоятельно оказывать медицинскую помощь, они всё еще нуждаются в контроле со стороны человека для гарантии действительности предоставляемой информации. Тем не менее они представляют собой важный шаг в направлении цифровизации и улучшения доступности медицинской помощи. Важно продолжать развитие и совершенствование этих систем, тем самым увеличивая их точность и безопасность. Это позволит сделать чат-боты полноценными инструментами как для пациентов, так и для медицинских работников.

1.2. РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ

Диалог пациента с врачом, как правило, не имеет какой-то четкой структуры. Поэтому главной задачей в разработке ассистента является извлечение необходимых сведений из диалогов в формате естественного языка. Для решения этой задачи возможно применение метода распознавания именованных сущностей – *Named Entity Recognition* (NER). С помощью данного метода [15] можно реализовать поиск, выделение и классификацию ключевых фрагментов текста, относящихся к медицинским терминам, например:

1) симптомы – сонливость, боль в горле, головокружение;
2) название анализов и их значений – ТТГ, ОАМ;
3) диагнозы – сахарный диабет, гипотиреоз.

Основной способ реализации NER – это методы машинного обучения, натренированные на размеченных медицинских текстах. Из моделей машинного обучения широкое распространение получили CRF (Conditional Random Fields – условное случайное поле) и LSTM (Long Short-Term Memory – длительная кратковременная память), однако модели на основе трансформеров, например BERT, способны демонстрировать более высокую точность, свыше 80 % [16], благодаря способности учитывать контекст предложения.

2. ОПИСАНИЕ ПОДХОДА

Основной задачей разрабатываемой системы является сбор ключевых данных о пациенте, обработка и анализ полученных сведений и постановка предварительного диагноза с перечнем необходимых анализов. Предварительный диагноз (или варианты диагнозов) может быть представлен только врачу, но скрыт от пациента. Выходной информацией, получаемой пациентом, является перечень необходимых анализов, результаты которых необходимо представить на очном приеме у врача-эндокринолога.

В ходе работы от медицинских специалистов в области эндокринологии были получены данные, представляющие собой перечень распространенных заболеваний, встречающихся в их практике, таких как сахарный диабет, гипотиреоз, гипертиреоз, феохромоцитома, синдром Конна, синдром Кушинга, надпочечниковая недостаточность. Данные включают описание наиболее характерных симптомов и жалоб для конкретного заболевания, а также наводящие вопросы, помогающие в диагностировании болезни. Для каждого заболевания указан необходимый перечень лабораторных и клинических исследований перед посещением профильного специалиста.

Разрабатываемая система построена по модульному принципу. Основу системы составляют три ключевых модуля, взаимодействующих друг с другом, для последовательной обработки информации и формирования диагностического заключения. Первый модуль системы отвечает за обработку текстовых жалоб пользователя с помощью модели извлечения симптомов. Ключевой особенностью этого модуля является его способность работать с разнообразными вариантами описания симптомов, включая разговорные и неформальные выражения. Это достигается благодаря семантической нормализации, которая преобразует разнообразные формулировки симптомов в стандартизированные медицинские категории. Выделенные симптомы передаются в модуль классификации диагнозов. Полученные данные анализируются, и на основе этого вычисляется вероятность наличия определенного заболевания. Если уверенность в диагнозе оказывается недостаточно высокой, система переходит в интерактивный режим, в котором задает пользователю уточняющие вопросы. Диалоговый модуль анализирует текущее состояние классификации и определяет, дополнительная информация о каком симптоме окажет наибольшее влияние на уточнение результатов. Вопросы к пользователю формулируются на естественном языке по заранее заданным шаблонам. Ответы пользователя обрабатываются с учетом возможных синонимичных выражений. Все модули интегрированы в единый сервисный слой, который координирует их взаимодействие, управляет потоком данных и формирует итоговый ответ для пользователя. Такая архитектура обеспечивает не только точность диагностики,

но и удобство использования. Разделение системы на модули позволяет при необходимости без высоких затрат корректировать модели и классификаторы без глобальных изменений всей архитектуры.

2.1. БАЗОВЫЙ АЛГОРИТМ ФУНКЦИОНИРОВАНИЯ ДИАЛОГОВОЙ СИСТЕМЫ

Процесс распознавания эндокринологических заболеваний можно разделить на три этапа (рис. 1): сбор анамнеза, обработка входных данных и классификация диагноза. В свободной форме пациент вводит жалобы, в которых описывает, что его беспокоит.

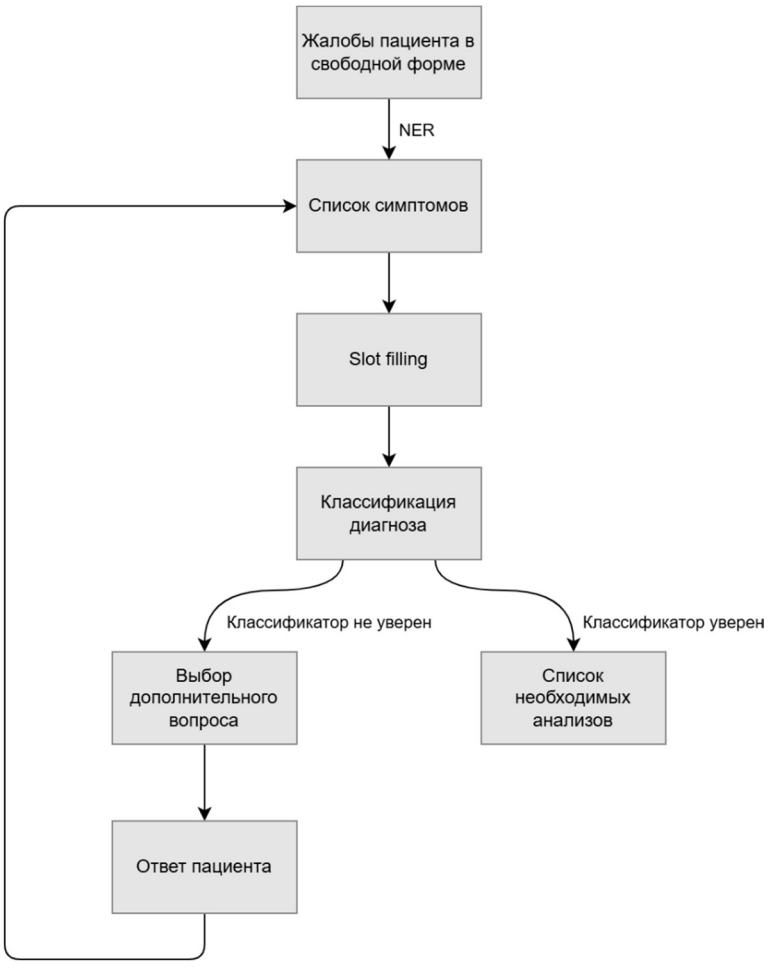


Рис. 1. Алгоритм работы системы

Fig. 1. System algorithm

Тексты уточняющих вопросов могут быть сгенерированы как заранее для каждого возможного симптома, так и во время диалога с подстановкой названия симптома в шаблоны вида «Вы испытывали {симптом} в последнее время?», «Отмечаете ли вы {симптом}?» и т. д. Однако важным аспектом является

генерация вопросов в закрытой форме, ответ на которую предполагает только два варианта – «да» или «нет». Такой подход призван снизить вычислительную нагрузку на систему поиска именованных сущностей и упростить последующую обработку ответов пользователя. Результаты ответа пользователя интерпретируются таким образом, что при получении ответа «да» признак наличия соответствующего симптома фиксируется положительным значением (единицей), при получении ответа «нет» – отрицательным значением (минус единицей). Эти значения записываются в специальную структуру, представляющую собой вектор симптомов, где каждому потенциальному симптому всегда соответствует определенное состояние: «точно имеется», «точно отсутствует» или «неизвестно».

Этап обработки начинается с применения к полученному тексту модели NER, которая обучена определять и извлекать сущности, относящиеся к симптомам. Далее выполняется *slot filling* – процесс заполнения матрицы симптомов. По полученному списку симптомов модель классификации пытается определить наиболее вероятный диагноз. Если классификатор не уверен, ему не хватает данных, то система возвращается на этап сбора анамнеза и инициирует процесс выбора уточняющего вопроса с целью получить больше данных. После ответа пациента система определяет и обновляет наличие симптомов и заново передает данные в модель классификатора. Система продолжает процесс выбора уточняющего вопроса до тех пор, пока классификатор не будет уверен.

В результате работы, если классификатор уверен и ему хватило входных данных, программа выдает возможный диагноз и список необходимых анализов, которые стоит сдать для подтверждения или опровержения поставленного диагноза. Для повышения точности определения диагноза необходимо предусмотреть механизм формирования дополнительных вопросов, который должен активироваться в случае недостаточной уверенности классификатора по результатам первичной обработки жалоб пациента в свободной форме. Основной целью такого механизма является получение дополнительной информации от пользователя для снижения неопределенности диагноза.

В качестве дополнительных вопросов подразумеваются уточняющие запросы о наличии или отсутствии симптомов, наиболее критичных для дифференцирования между несколькими потенциальными диагнозами. Поэтому для выбора вопросов следует руководствоваться принципом максимизации прироста информации и для выбора вопроса выбирать те симптомы, наличие или отсутствие которых наиболее существенно влияет на распределение вероятностей диагнозов. Таким образом предпочтение будет отдаваться симптомам, высоко специфичным для одного из возможных диагнозов, благодаря чему будет достигнут критерий минимизации числа вопросов.

Дополнительно, чтобы не перегружать пользователя, имеет смысл ограничить число задаваемых вопросов. Если модель классификации не смогла определить заболевание за ограниченный круг вопросов, то это может сигнализировать о том, что либо заболевание пользователя не имеет отношения к эндокринологии, либо случай пользователя слишком специфичный. В любом из двух возможных случаев, если модель долго остается неуверенной, лучше перенаправить пользователя к специалисту.

2.2. РЕАЛИЗАЦИЯ МОДУЛЯ ИЗВЛЕЧЕНИЯ СИМПТОМОВ

Модуль извлечения симптомов из текста является основным компонентом системы, так как именно он отвечает за обработку жалоб пациентов и выделение из них отдельных сущностей, связанных с симптомами. Основная задача этого модуля заключается в преобразовании неструктурированных текстовых данных, введенных пользователем, в структурированные данные в формате, пригодном для дальнейшего использования. Для решения этой задачи используются методы обработки естественного языка, включающие извлечение именованных сущностей и семантическую нормализацию.

Разработанный модуль включает в себя несколько этапов: подготовку данных, работу модели NER и нормализацию найденных симптомов. Подробно описываются этапы реализации модуля, такие как выбор и подготовка набора данных, обучение модели, оценка ее эффективности и методы нормализации симптомов.

Для обучения модели NER был взят открытый набор данных пользовательских отзывов о лекарственных препаратах на русском языке – *The Russian Drug Reaction Corpus* (RuDReC) [17], который предназначен для извлечения информации в области здравоохранения для обнаружения побочных эффектов. В настоящий момент набор данных состоит примерно из 5000 размеченных отзывов, собранных из различных интернет-источников. Из этого количества записей более половины содержат сущности, связанные с лекарствами и заболеваниями. Этот набор данных был выбран в работе в качестве основного, так как он является одним из немногих доступных открытых специализированных наборов данных на русском языке, содержащих качественную экспертную разметку медицинских сущностей. Также этот набор данных предоставляет достаточно репрезентативную выборку для дообучения языковых моделей.

Классы аннотированных сущностей в наборе данных представляют разнообразные аспекты, связанные с лекарствами и их воздействием, а также имеют дополнительную разметку связей на уровне выражений для идентификации тонких подтипов. Однако основные типы сущностей включают следующие классы.

1. Drug Name – упоминание наименований препаратов и активных веществ.
2. Drug Class – упоминание классов или предназначений лекарственных препаратов, как, например, «противовоспалительные», «иммуномодулирующие» и т. д.
3. Drug Form – форма выпуска лекарственного препарата.
4. Adverse Drug Reaction (ADR) – описанные пользователем побочные эффекты, связанные с приемом лекарств.
5. Drug Indication (DI) – показания к применению препарата, включающие в себя симптомы или упоминания заболеваний.
6. Findings – описанные пользователем упоминания изменений состояния здоровья, включающие в себя описания ухудшения или улучшения после приема препарата.

В рамках настоящей работы наибольшую ценность представляют последние три пункта, поскольку именно они содержат информацию о симптомах пользователей. По этой причине текстовые записи из набора данных были

отфильтрованы по критерию наличия аннотированных сущностей интересующих типов. Если запись содержала хотя бы одну сущность из классов ADR, DI или Findings, то она добавлялась в результирующий набор данных, при этом информация о разметке сущностей нецелевых типов в такой записи игнорировалась. Дополнительно отфильтрованный набор данных был дополнен записями, которые не содержат ни одной именованной сущности, чтобы эти записи составляли не более 10 % от общего размера набора данных. В результате размер итогового набора данных составил 1550 записей.

Для генерации синтетического набора данных использовалась большая языковая модель (БЯМ) *DeepSeek* [18]. В эту модель подавались подобранные эмпирическим путем задания на генерацию, составленные таким образом, чтобы модель перефразировала текст и/или заменяла упоминания симптомов в записях из RuDReC. Также модели выдавались задания на генерацию новых записей по реальным примерам из набора данных. Опытным путем было выявлено, что БЯМ плохо определяет позиции символов в тексте и не способна выдавать разметку в виде пары значений start и end. Для решения этой проблемы тексты заданий для модели были скорректированы таким образом, чтобы вместо позиций начала и конца именованной сущности модель выдавала в разметке вырезанный фрагмент текста сущности. Затем по этим фрагментам в исходной строке алгоритмическими методами определялись начальная и конечная позиция.

Также несмотря на то, что БЯМ хорошо справлялась с распознаванием упоминаний симптомов, у модели наблюдались проблемы с автоматической классификацией размеченных ею сущностей при отнесении симптомов к корректной категории. Отчасти это связано с отсутствием унифицированного стандарта наименований классов, из-за чего одни и те же симптомы могли быть отнесены к разным категориям. По этой причине наравне с тем, что и в RuDReC разметка не имеет указаний на конкретные типы симптомов, в разрабатываемой системе модель NER должна будет предсказывать только один обобщенный класс, сигнализирующий о том, что текстовая последовательность является описанием симптома, а дальнейшее уточнение его типа будет выполняться на последующих этапах обработки с использованием других методов. С помощью такого подхода удалось в короткие сроки сгенерировать набор синтетических данных, объем которого составил более 100 записей, что является достаточным для представления до модели NER информации о симптомах, специфических для эндокринологии. В совокупности сочетание реального набора данных с синтетически сгенерированными примерами позволит компенсировать нехватку релевантных в целевой предметной области данных и обеспечить большую адаптацию модели к специфике задачи, предоставив достаточно большой объем обучающих данных.

Для решения задачи NER была выбрана предварительно обученная, в том числе и на русскоязычных текстах, модель, основанная на архитектуре BERT cointegrated/LaBSE-en-ru. Исходный набор данных был разбит на две выборки: тренировочную и валидационную в соотношении 80 : 20 соответственно. Перед подачей данных в модель необходимо было провести их токенизацию. Исходный текст записи разбивался на подслова с помощью токенизатора, предоставляемого вместе с выбранной моделью, и дополнялся специальными токенами так, чтобы выровнять и привести все записи к единому размеру

в 256 токенов. Целевые метки сущностей также необходимо было привести в формат токенов с использованием BIO-нотации. Основная сложность заключается в том, что токенизатор может разбить одно слово из сущности на несколько подслов, в таком случае невозможно просто сопоставить номер слова из сущности с номером токена. Так как токенизатор предоставляет индекс начала и конца каждого токена в исходной строке, для преобразования меток из формата (start, end) в BIO-формат был использован следующий алгоритм.

1. Производится итерация по каждому интервалу из начала и конца токенов.

2. Осуществляется пропуск сущностей из разметки, конец которых меньше начала текущего интервала.

3. Если начало текущего интервала больше или равно началу сущности и конец интервала меньше или равен концу сущности из разметки, то проверяется предыдущий интервал:

а) если предыдущего интервала нет либо он не пересекается с текущей сущностью, то токену присваивается метка «B» начала сущности;

б) в ином случае токену присваивается метка «I» продолжения сущности.

4. Если интервал токена не пересекается с разметкой сущности, то такому токену присваивается метка «O».

С помощью описанного алгоритма получается разметка, в которой каждому токену из входящей последовательности сопоставлена целевая метка, которую будет пытаться предсказать модель. В цикле обучения на вход модели подается пакет токенизированных текстовых последовательностей. Для каждого входного токена модель предсказывает метку, вычисляется ошибка модели и с помощью оптимизатора корректируются веса модели. Обучение модели проводилось с помощью сервиса Google Colab, который представляет собой облачную среду для разработки и выполнения кода на языке Python. Также данный сервис предоставляет доступ к видеокартам Nvidia T4, которые, обладая видеопамятью в 24 Гб, являются незаменимыми при обучении языковых моделей. Гиперпараметры для обучения были подобраны эмпирическим путем. Основными параметрами при обучении стали оптимизатор AdamW с параметром learning_rate = 2e-5 и размером пакета 32. Обучение проводилось на 20 эпохах, после которых модель выходила на стабильное плато и переставала улучшать показатели.

При использовании модели в рабочей среде для получения предсказания необходимо выполнить аналогичную токенизацию текста и подать его на вход модели, затем по предсказанным меткам восстановить начало и конец предсказанных сущностей. В дополнение к предсказанию меток модель выдает некоторую меру уверенности в своем ответе, которая может быть использована для отсечения по заданному пороговому значению предсказаний, в которых модель слабо уверена.

2.3. МЕТРИКИ КАЧЕСТВА

Основными метриками при обучении моделей NER являются *F1-Score*, *Precision* и *Recall*. Метрика *Precision (PR)* характеризует, насколько уверенно

модель предсказывает сущности, т. е. сколько из всех предсказанных моделью сущностей было предсказано верно, и рассчитывается по формуле

$$PR = \frac{TP}{TP + FP},$$

где TP – количество истинно-положительных предсказаний, FP – количество ложноположительных предсказаний.

Метрика *Recall* (*RC*) является метрикой полноты и показывает, насколько внимательно модель находит сущности, т. е. сколько из всех реальных сущностей она нашла, и рассчитывается по формуле

$$RC = \frac{TP}{TP + FN},$$

где FN – количество ложнонегативных предсказаний.

Метрика *F1-Score (F1S)* представляет собой гармоническое среднее между точностью и полнотой и рассчитывается по формуле

$$F1S = 2\frac{PR \cdot RC}{PR + RC}.$$

В рамках данной задачи интерес представляют высокие показатели точности и в меньшей степени полнота, так как если будет пропущена какая-либо сущность, это не так критично, потому что она потом будет доопределена на последующих этапах общения с пользователем. Однако сильный дисбаланс между этими метриками является негативным маркером того, что модель обучилась плохо, поэтому метрика *F1-Score* также важна как показатель баланса между точностью и полнотой модели. Как правило, признаком дисбаланса между *Precision* и *Recall* является существенное различие значений этих метрик, при этом *F1-Score* заметно ниже обоих показателей.

В результате обучения модели были достигнуты следующие показатели метрик: $F1S$ = 0,972 и PR = 0,971. Однако полученные метрики не в полной мере отображают действительную способность модели предсказывать метки именованных сущностей, так как наблюдается колоссальный дисбаланс, смещенный в сторону метки «О», который связан с тем, что в естественных данных большая часть слов не является целевыми медицинскими именованными сущностями. Таким образом, эти метрики в большей степени показывают способность модели определять, что слово не является именованной сущностью. Поэтому для более объективной оценки использовались метрики $F1S_{clear}$ и PR_{clear}. Для их подсчета не учитывались пары, в которых целевой меткой была метрика «О», и модель предсказала метку «О». Эти метрики обозначены индексом *clear* и вычисляются по следующим формулам:

$$TP_{clear} = \sum_{i=1}^{N}\left[y_{true}^{(i)} \neq O \wedge y_{pred}^{(i)} \neq O \wedge y_{true}^{(i)} = y_{pred}^{(i)} \right],$$

где $y_{true}^{(i)}$ – истинное значение целевой переменной, $y_{pred}^{(i)}$ – спрогнозированное значение целевой переменной, $\wedge$ – логическое «И», O – метка «О»;

$$FP_{clear} = \sum_{i=1}^{N}\left[\left(y_{true}^{(i)} \neq O \wedge y_{pred}^{(i)} \neq O \wedge y_{true}^{(i)} = y_{pred}^{(i)}\right) \vee \left(y_{true}^{(i)} = O \wedge y_{pred}^{(i)} \neq O\right)\right],$$

где $\vee$ – логическое «ИЛИ»;

$$FN_{clear} = \sum_{i=1}^{N}\left[y_{true}^{(i)} \neq O \wedge y_{pred}^{(i)} = O\right],$$

$$PR_{clear} = \frac{TP_{clear}}{TP_{clear} + FP_{clear}},$$

$$RC_{clear} = \frac{TP_{clear}}{TP_{clear} + FN_{clear}},$$

$$F1S_{clear} = 2\frac{PR_{clear} \cdot RC_{clear}}{PR_{clear} + RC_{clear}}.$$

Эти метрики помогают более достоверно оценить качество распознавания именно сущностей и показывают, как модель справляется только с именованными сущностями, без искажения из-за дисбаланса меток «O». Для обученной модели метрики составили: $F1S_{clear} = 0{,}951$ и $PR_{clear} = 0{,}950$.

Также для наглядной визуализации оценки обученной модели были построены матрицы ошибок (*confusion matrix*) в нескольких вариантах. Матрица ошибок широко используется для оценки качества классификации в задачах машинного обучения и представляет собой квадратную матрицу, элементы которой отражают частоту соответствия между истинными и предсказанными метками классов. Строки матрицы соответствуют истинному классу, а столбцы – предсказанному моделью, при этом элементы на главной диагонали характеризуют корректно классифицированные объекты, а остальные элементы описывают случаи ошибочной классификации.

На рис. 2 представлена стандартная матрица ошибок, где видно преобладание диагональных элементов для класса «O», что подтверждает его доминирование в данных. Для того чтобы более детально показать качество распознавания именно целевых медицинских сущностей, была построена сфокусированная матрица, в которой аналогично подсчету метрик выше были исключены случаи верно предсказанных метрик «O».

На рис. 3 представлена сфокусированная версия матрицы ошибок, где видно, что модель с достаточной точностью выделяет медицинские сущности из текста с минимальными перекрестными ошибками между метками начала и продолжения сущности.

Дополнительно для комплексной оценки производительности модели NER целесообразно проанализировать не только раздельные метки в BIO-формате, но и агрегированный результат, в котором метки начала и продолжения сущности объединены в общий класс симптома *SYM*. Такая агрегированная матрица (рис. 4) позволяет оценить, насколько корректно модель идентифицирует сущности в целом, так как при использовании модели в режиме предсказания некоторые ошибки в разметке начальных и внутренних меток возможно исправить алгоритмическими методами.

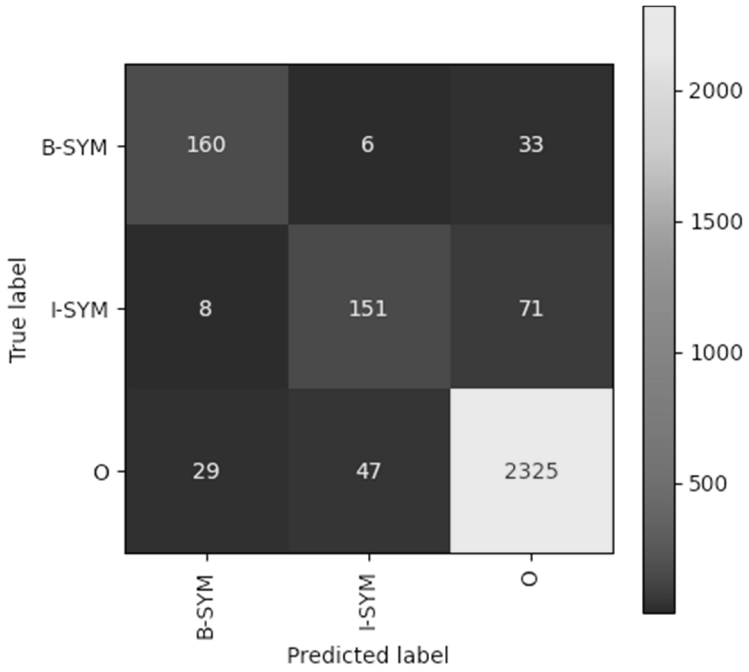


Рис. 2. Матрица ошибок с учетом класса «O»

Fig. 2. The error matrix taking into account class “O”

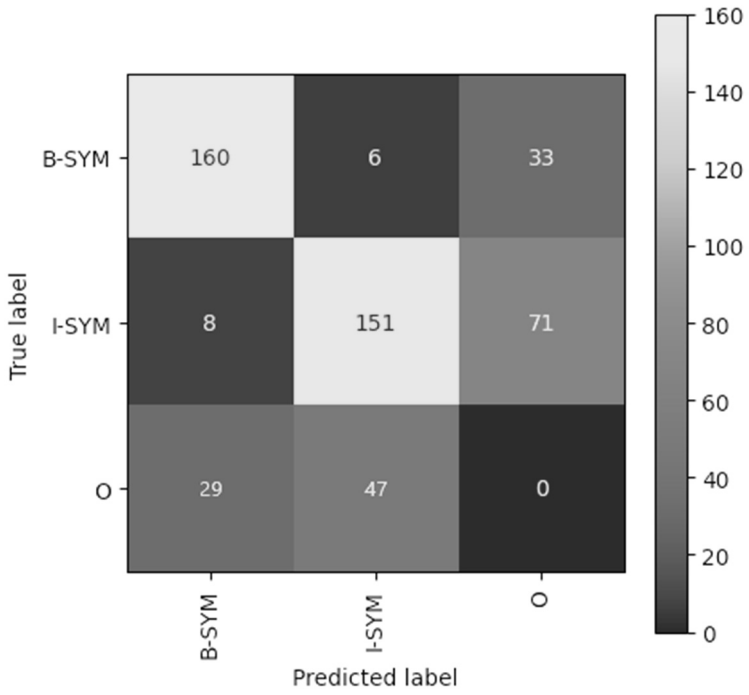


Рис. 3. Сфокусированная матрица ошибок

Fig. 3. The focused error matrix

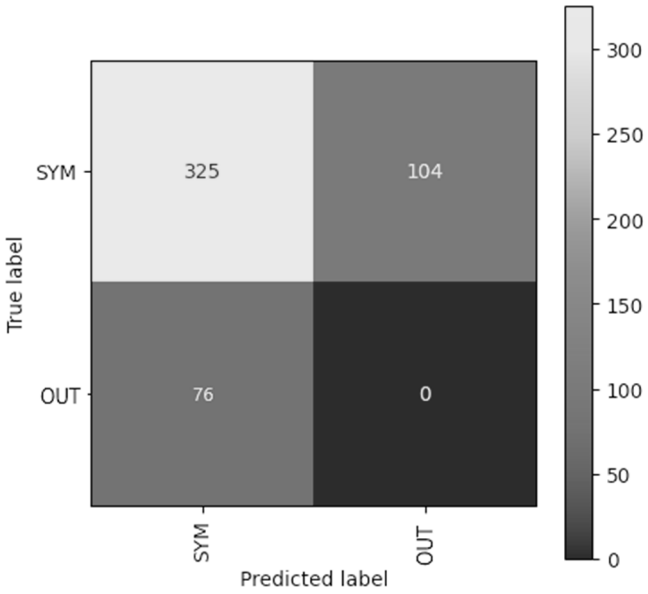


Рис. 4. Агрегированная матрица ошибок

Fig. 4. The aggregated error matrix

Таким образом, приведенные метрики качества и их визуализация в виде матриц ошибок подтверждают эффективность обученной модели NER в задаче распознавания симптомов пациентов.

2.4. НОРМАЛИЗАЦИЯ СИМПТОМОВ

Несмотря на достаточное качество распознавания сущностей описанной выше моделью, она предсказывает только факт наличия симптома, но не определяет его конкретный класс. Поэтому распознанные моделью симптомы требуют дальнейшей нормализации для отнесения их к конкретным медицинским категориям. Для решения этой задачи был разработан и реализован дополнительный модуль нормализации симптомов. Для приведения симптомов к стандартизированному виду используется метод семантической нормализации. Эталонные классы симптомов представлены в виде набора синонимичных формулировок, что обеспечивает адаптивность к разнообразию формулировок симптомов пациентами. Алгоритм нормализации использует предобученную модель LaBSE-en-ru для перевода эталонных формулировок и текста предсказанной сущности в векторные представления. Далее вектор предсказанного текста поочередно сравнивается с каждым вектором из набора эталонных примеров для каждого возможного класса симптомов, используя меру косинусного сходства *similarity* (SMT), вычисляемую по формуле [19]

$$SMT(A,\ B) = \frac{AB}{\| A \| \| B \|} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i{}^2}\ \sqrt{\sum_{i=1}^{N} b_i{}^2}},$$

где A и B – векторы текста сущности и эталонного примера соответственно.

Косинусное сходство позволяет оценить близость понятий, текст преобразуется в векторное пространство, такое что близкие по смыслу фразы имеют схожие направления векторов, даже если они выражены разными словами. Также возможно сравнение векторов с использованием евклидова расстояния, однако мера косинусного сходства не зависит от длины самих векторов. Инвариантность к длине текста важна при сравнении коротких названий симптомов с их развернутыми описаниями.

После вычисления косинусных расстояний вычисляется максимальная близость для каждого класса. Если близость превышает заданное пороговое значение, то симптом относится к соответствующему классу, в противном случае помечается как неизвестный. Более формализованное описание данного алгоритма включает в себя заданное входное множество классов симптомов:

$$C = \{c_1,\ c_2,...,\ c_n\}\ .$$

Для каждого класса симптомов задается множество синонимических наиболее распространенных названий этого симптома и разговорных формулировок его проявления:

$$E_c = \{e_1,\ e_2,...,\ e_n\}\ .$$

Векторное представление текста t в эмбеддинг v_t с помощью модели M задается формулой

$$v_t = M(t)\ .$$

Тогда для входного текста распознанной сущности t класс симптома вычисляется следующей формулой:

$$c'(t) = \arg\max_{c \in C} \left(\max_{e \in E_c} SMT(M(t),\ M(e)) \right),\ \text{если}\ \max \geq \theta,$$

иначе *unknown*.

Здесь c' – предсказанный класс симптома; t – входной текст симптома, распознанного моделью NER; C – множество классов симптомов; E_c – множество примеров для конкретного класса симптомов; M – модель векторизации; θ – порог уверенности.

Из формулы следует, что если близость для всех классов ниже некоторого порогового значения θ, то симптом помечается как *unknown*. Такая фильтрация обеспечивает возврат результата только при высокой степени схожести, что снижает количество ложных срабатываний и помогает отбросить симптомы, не имеющие отношения к эндокринологии. Подобные симптомы не передаются дальше в модель классификации. Таким образом, входной текст классифицируется по ближайшему классу на основе анализа векторных представлений, что позволяет учитывать разнообразие формулировок пациентов. Благодаря этому нормализатор преобразует разрозненные описания жалоб пациентов в стандартизированные классы, что позволяет перейти от распознавания факта упоминания симптома к его содержательной интерпретации, необходимой для последующего анализа и предсказания предварительного диагноза.

3. КЛАССИФИКАЦИЯ ЭНДОКРИНОЛОГИЧЕСКОГО ДИАГНОЗА

Для определения предварительного диагноза пациента по его жалобам был реализован модуль классификации по списку выявленных симптомов. В рамках настоящей работы были выбраны два алгоритма классификации: логистическая регрессия и наивный байесовский классификатор. Такой выбор обусловлен эффективностью этих алгоритмов в задачах с ограниченным объемом данных и интерпретируемостью их результатов. Логистическая регрессия способна моделировать вероятности принадлежности к классам через сигмоидальную функцию:

$$(y = 1 \vee x) = \frac{1}{1 + e^{-(wTx+b)}},$$

где w – веса признаков, T – операция транспонирования, b – смещение, x – вектор симптомов.

Обучение модели логистической регрессии [20] сводится к нахождению параметров w и b, минимизирующих функцию потерь L:

$$L(w,\ b) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(P(y_i = 1 \vee x_i)\right) + (1 - y_i)\log\left(1 - P(y_i = 1 \vee x_i)\right)\right],$$

где N – количество примеров в выборке, y_i – истинная метка класса для i-го примера.

Подбор параметров w и b выполняется методом градиентного спуска на протяжении заданного числа итераций:

$$w^{t+1} = w^t - \eta \nabla_w L,$$

$$b^{t+1} = b^t - \eta \frac{\partial L}{\partial b},$$

где η – скорость обучения, t – номер итерации, η – скорость обучения, $\nabla_w L$ – градиент функции потерь L по весам w, $\frac{\partial L}{\partial b}$ – частная производная функции по смещению b.

Наивный байесовский классификатор [21] основывается на теореме Байеса с предположением о независимости признаков:

$$P(y \vee x) = \frac{P(y)\prod_{i=1}^{n} P(x_i \vee y)}{P(x)}.$$

Априорная вероятность $P(y)$ вычисляется как доля примеров класса y в обучающей выборке:

$$P(y) = \frac{NCL_y}{N},$$

где NCL_y – количество примеров класса y.

Условные вероятности $P(x_i \vee y)$ для категориальных признаков вычисляются как

$$P(x_i \vee y) = \frac{NCL_{y(x_i)}}{NCL_y},$$

где $NCL_{y(xi)}$ – количество примеров класса y с признаками x_i.

Исходные данные для обучения алгоритмов были преобразованы в файл формата CSV, содержащего бинарные признаки наличия симптомов и текстовые названия возможных диагнозов. Признаки в файле закодированы таким образом, что значение «1» обозначат наличие симптома, а значение «–1» – отсутствие симптома. Текстовые названия диагнозов являются целевой переменной, которую должны предсказывать выбранные алгоритмы. Однако в силу возможности их работы только с числовыми переменными необходимо применение метода LabelEncoder, который преобразует текстовые метки в числовые и хранит в себе список соответствия. Оба алгоритма были обучены на одном наборе данных, разделенных на признаки и целевую переменную. Гиперпараметры моделей, такие как регуляризация для логистической регрессии или априорные вероятности для байесовского классификатора, использовались со значениями по умолчанию, так как объем данных не требовал сложной настройки. Обучение проводилось на полном наборе данных без выделения отдельной тестовой выборки из-за его ограниченного размера.

Модели после обучения были сохранены в виде объекта, готового к использованию в реальном времени. При поступлении списка симптомов, выявленных у пользователя, они преобразуются в вектор признаков, согласованный с порядком колонок в тренировочных данных. При этом значение «1» соответствует установленному у пациента наличию симптома, значение «–1» – отсутствию симптома, а пропущенные значения заполняются значением «0», показывающим, что информация о симптоме не предоставлена. Это позволяет модели обрабатывать случаи, когда информация о симптоме еще не была получена от пользователя. Диалог с пользователем реализован в виде консольного приложения. В результате модуль возвращает вероятности принадлежности набора симптомов к каждому возможному диагнозу, что позволяет ранжировать результаты и определить наиболее вероятный вариант. Настройки системы позволяют отображать только рекомендуемые для сдачи анализы, а предполагаемый диагноз демонстрировать врачу. В случае недостаточной уверенности модуля в наиболее вероятном варианте необходима дальнейшая его проработка и получение от пользователя дополнительной информации о наличии или отсутствии симптомов, информация о которых еще не известна.

ЗАКЛЮЧЕНИЕ

Представленная диалоговая система предназначена для поддержки принятия врачебных решений и упрощения процесса приема пациентов в условиях недостатка квалифицированных специалистов. В текущей версии система охватывает классификацию наиболее распространенных эндокринологических

заболеваний. Расширение этого списка требует сбора дополнительных данных. Большинство популярных рассмотренных диалоговых систем являются зарубежными разработками и обладают широким перечнем функций: от предоставления справочной информации по общим медицинским вопросам и рекомендаций по выбору специалиста до осуществления записи к врачу. Особенностью предложенной системы является привязка к конкретной отрасли медицины. В отличие от многих чат-ботов, ориентированных преимущественно на удобство пациентов, предложенная разработка позволяет снизить нагрузку на медицинских специалистов за счет сбора анамнеза и обработки полученных результатов. Основным недостатком системы является обучение на наборах данных малого объема, которые содержат ограниченное количество примеров, специфичных для эндокринологии, что может снижать точность редких или сложных случаев. Также система может испытывать трудности с пониманием неоднозначных формулировок жалоб.

Модель обработки естественного языка ориентирована на обработку только русскоязычных текстов, что может ограничивать ее применение в международных масштабах. Однако именно обработка только русскоязычных текстов делает систему уникальной среди популярных чат-ботов, осуществляющих диалог только на английском языке. Таким образом, на данный момент система не заменяет профессиональную медицинскую диагностику, а лишь оптимизирует процесс сбора первичной информации, предоставляя предварительные диагнозы, которые требуют обязательной проверки врачом.

Несмотря на текущие ограничения, система обладает значительным потенциалом для развития и внедрения в практическое здравоохранение. Дальнейшие исследования и доработки позволят повысить ее точность, универсальность и полезность для пациентов и врачей. Основными направлениями дальнейших исследований являются работы по расширению обучающих наборов данных, а также применение системы для других областей медицины.

Благодарности. Авторы выражают благодарность профессору кафедры вычислительной техники, доктору медицинских наук Мелдо Анне Александровне за консультирование в проведении исследования и разработки.

СПИСОК ЛИТЕРАТУРЫ

1. *Терещенко И.В.* Пропедевтика эндокринных заболеваний (методы исследования больных с заболеваниями эндокринной системы): монография. – Чебоксары: Среда, 2022. – 104 с. – ISBN 978-5-907411-89-0. – DOI: 10.31483/a-10348.

2. *Тимофеев Е.В., Парфенова Н.Н., Реева С.В.* Обследование эндокринной и кроветворной систем в клинике внутренних болезней: учебное пособие. – СПб.: Изд-во С.-Петерб. гос. педиатр. мед. ун-та, 2022. – 32 с.

3. *Медведева Е.И., Александрова О.А., Крошилин С.В.* Телемедицина в современных условиях: отношение социума и вектор развития // Экономические и социальные перемены: факты, тенденции, прогноз. – 2022. – Т. 15, № 3. – С. 200–222. – DOI: 10.15838/esc.2022.3.81.11.

4. *Баранов А.А, Вишнева Е.А., Намазова-Баранова Л.С.* Телемедицина – перспективы и трудности перед новым этапом развития // Педиатрическая фармакология. – 2013. – Т. 10, № 3. – С. 6–11. – DOI: 10.15690/pf.v10i3.691.

5. *Хайруллин А.М., Зарипова Р.С.* Использование чат-ботов и онлайн-консультации как будущее медицины // Информационные технологии в строительных, социальных и экономических системах. – 2020. – № 3 (21). – С. 74–76.

6. *Аксёнова Е.И., Медведева Е.И., Крошилин С.В.* Чат-боты – современная реальность консультирования в медицине // Здравоохранение Российской Федерации. – 2023. – Т. 67, № 5. – С. 403–410. – DOI: 10.47470/0044-197X-2023-67-5-403-410.

7. Babylon. – URL: https:// www.babylonhealth.com (accessed: 04.02.2025).

8. Ada Health. – URL: https://ada.com (accessed: 04.02.2025).

9. ByoHealth. – URL: https://www.buoyhealth.com (accessed: 04.02.2025).

10. Florence. – URL: https://florencehc.com (accessed: 04.02.2025).

11. Symptomate. – URL: https://symptomate.com (accessed: 04.02.2025).

12. Sensely. – URL: https://sensely.com (accessed 04.02.2025).

13. *McMahon G.T.* The risks and challenges of artificial intelligence in endocrinology // The Journal of Clinical Endocrinology & Metabolism. – 2024. – Vol. 109 (6). – P. e1468-e1471.

14. Roles, users, benefits, and limitations of chatbots in health care: rapid review / M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, B. Lebouche // Journal of Medical Internet Research. – 2024. – Vol. 26. – P. e56930.

15. Feature extraction method from electronic health records in Russia / D. Gavrilov, A. Gusev, I. Korsakov, R. Novitskiy, L. Serova // Proceedings of the 26th Conference of FRUCT Association. – 2020. – P. 497–500. – ISBN 978-952-69244-2-7.

16. The Russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews / E. Tutubalina, I. Alimova, Z. Miftahutdinov, A. Sakhovskiy, V. Malykh, S. Nikolenko // Bioinformatics. – 2020. – Vol. 36 (2). – P. 243–250. – DOI: 10.1093/bioinformatics/btaa675.

17. Exploring DeepSeek: A survey on advances, applications, challenges and future directions / Z. Deng, W. Ma, Q.-L. Han, W. Zhou, X. Zhu, S. Wen, Y. Xiang // IEEE/CAA Journal of Automatica Sinica. – 2025. – Vol. 12 (5). – P. 872–893.

18. *Chopra S., Hadsell R., LeCun Y.* Learning a similarity metric discriminatively, with application to face verification // IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). – 2005. – Vol. 1 – P. 539–546.

19. *Han J., Morag C.* The influence of the sigmoid function parameters on the speed of backpropagation learning // From natural to artificial neural computation / ed. by J. Mira, F. Sandoval. – Springer, 1995. – P. 195–201. – (Lecture Notes in Computer Science; vol. 930).

20. *Maron M.E.* Automatic indexing: An experimental inquiry // Journal of the ACM. – 1961. – Vol. 8 (3). – P. 404–417.

Мамикьян Маргарита Сергеевна, студентка кафедры вычислительной техники Санкт-Петербургского государственного электротехнического университета «ЛЭТИ». Основное направление научных исследований – машинное обучение, системный анализ, информационные системы. E-mail: mmamikyan2001@gmail.com

Бекенева Яна Андреевна, кандидат технических наук, доцент кафедры вычислительной техники Санкт-Петербургского государственного электротехнического университета «ЛЭТИ». Основное направление научных исследований – методы обработки данных, машинное обучение, информационные системы. Имеет более 50 печатных работ. E-mail: yabekeneva@etu.ru

Mamikyan Margarita S., student at the Department of Computer Science and Engineering, Saint Petersburg Electrotechnical University “LETI”. Her research interests are currently focused on machine learning, system analysis, and information systems. E-mail: mmamikyan2001@gmail.com

Bekeneva Yana A., PhD (Eng.), associate professor at the Department of Computer Science and Engineering, Saint Petersburg Electrotechnical University “LETI”. Her research interests are currently focused on data processing, machine learning, and information systems. She has more than 50 publications. E-mail: yabekeneva@etu.ru

DOI: 10.17212/2782-2001-2026-3-75-96

Processing text data in medical dialogue systems using the example of a virtual assistant for an endocrinologist [*]

M.S. MAMIKYAN[a], *YA.A. BEKENEVA*[b]

Saint Petersburg Electrotechnical University "LETI", 5 Professora Popova Street, St-Petersburg, 197022, Russian Federation

[a] *mmamikyan2001@gmail.com* [b] *yabekeneva@etu.*

Abstract

The proportion of patients with endocrine pathologies is constantly increasing in all economically developed countries of the world. In Russia, over the past 6 years, the prevalence of endocrine diseases has increased by 26%, and primary morbidity rates by 36%. One of the main problems that aggravate the epidemiological situation is the shortage of local therapists and the uneven distribution of medical specialists between rural and urban regions. Residents of sparsely populated areas are forced to travel to neighboring cities to receive the necessary medical care, which entails additional time and financial costs. It is also worth noting that determining the necessary tests for delivery can take a lot of time, and rushing during the appointment can lead to the delivery of redundant tests. To optimize this process, it is important to determine in advance what tests the patient needs, which requires a preliminary diagnosis. A disease recognition system based on patient complaints can help with this task, which can automate the preliminary diagnosis and avoid unnecessary studies. The presented software solution is based on a combination of named entity recongition for extracting symptoms from free-form patient complaints and machine learning methods for determining a diagnosis. The RuDRec dataset was used for training, along with synthetic data. An algorithm for dialogue with the user to clarify symptoms and generate a list of necessary tests is described. Such a system will help reduce cases of self-medication, when patients, without precise recommendations, follow the advice of friends. Thus, the formation of directions for tests based on high-quality primary consultation of patients, especially in sparsely populated areas, is an urgent task. The practical significance of the work lies in optimizing the load on medical institutions, reducing the risks of self-medication and improving diagnostics in remote regions.

Keywords: virtual assistant of endocrinologist, medical data processing methods, large language models, chatbot, dialog interface, health complaint, diagnostics, endocrinological disease, telemedicine, named entities, intelligent machine learning system

REFERENCES

1. Tereshchenko I.V. *Propedevtika endokrinnykh zabolevanii (metody issledovaniya bol'nykh s zabolevaniyami endokrinnoi sistemy)* [Propaedeutics of endocrinisms (Methods of examination of patients with diseases of the endocrine system)] Cheboksary, Sreda Publ., 2022. 104 p. ISBN 978-5-907411-89-0. DOI: 10.31483/a-10348.

2. Timofeev E.V., Parfenova N.N., Reeva S.V. *Obsledovanie endokrinnoi i krovetvornoi sistem v klinike vnutrennikh boleznei* [Examination of the endocrine and hematopoietic systems in the clinic of internal diseases]. St. Petersburg, St. Petersburg State Pediatric Medical University Publ., 2022. 32 p.

3. Medvedeva E.I., Aleksandrova O.A., Kroshilin S.V. Telemeditsina v sovremennykh usloviyakh: otnoshenie sotsiuma i vektor razvitiya [Telemedicine in modern conditions: the attitude of society and the vector of development]. *Ekonomicheskie i sotsial'nye peremeny: fakty, tendentsii, prognoz = Economic and Social Changes: Facts, Trends, Forecast*, 2022, vol. 15, no. 3, pp. 200–222. DOI: 10.15838/esc.2022.3.81.11.

[*] *Received 06 June 2025.*

4. Baranov A.A, Vishneva E.A., Namazova-Baranova L.S. Telemeditsina – perspektivy i trudnosti pered novym etapom razvitiya [Telemedicine – prospects and difficulties before a new stage of development]. *Pediatricheskaya farmakologiya = Pediatric Pharmacology*, 2013, vol. 10, no. 3, pp. 6–11. DOI: 10.15690/pf.v10i3.691.

5. Khairullin A.M., Zaripova R.S. Ispol'zovanie chat-botov i onlain-konsul'tatsii kak budushchee meditsiny [Using chat bots and online consultation as the future of medicine]. *Informatsionnye tekhnologii v stroitel'nykh, sotsial'nykh i ekonomicheskikh sistemakh*, 2020, no. 3 (21), pp. 74–76. (In Russian).

6. Aksenova E.I., Medvedeva E.I., Kroshilin S.V. Chat-boty – sovremennaya real'nost' konsul'tirovaniya v meditsine [Chatbots is the modern reality of consulting in medicine]. *Zdravookhranenie Rossiiskoi Federatsii = Health Care of the Russian Federation*, 2023, vol. 67, no. 5, pp. 403–410. DOI: 10.47470/0044-197X-2023-67-5-403-410.

7. *Babylon*. Available at: https:// www.babylonhealth.com (accessed 04.02.2025).

8. *Ada Health*. Available at: https://ada.com (accessed 04.02.2025).

9. *ByoHealth*. Available at: https://www.buoyhealth.com (accessed 04.02.2025).

10. *Florence.* Available at://florencehc.com (accessed 04.02.2025).

11. *Symptomate*. Available at: https://symptomate.com (accessed 04.02.2025).

12. *Sensely*. Available at: https://sensely.com (accessed 04.02.2025).

13. McMahon G.T. The risks and challenges of artificial intelligence in endocrinology. *The Journal of Clinical Endocrinology & Metabolism*, 2024, vol. 109 (6), p. e1468-e1471.

14. Laymouna M., Ma Y., Lessard D., Schuster T., Engler K., Lebouche B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. *Journal of Medical Internet Research*, 2024, vol. 26, p. e56930.

15. Gavrilov D., Gusev A., Korsakov I., Novitskiy R., Serova L. Feature extraction method from electronic health records in Russia. *Proceedings of the 26th Conference of FRUCT Association*, 2020, pp. 497–500. ISBN 978-952-69244-2-7.

16. Tutubalina E., Alimova I., Miftahutdinov Z., Sakhovskiy A., Malykh V., Nikolenko S. The Russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, 2020, vol. 36 (2), pp. 243–250. DOI: 10.1093/bioinformatics/btaa675.

17. Deng Z., Ma W., Han Q.-L., Zhou W., Zhu X., Wen S., Xiang Y. Exploring DeepSeek: A survey on advances, applications, challenges and future directions. *IEEE/CAA Journal of Automatica Sinica*, 2025, vol. 12 (5), pp. 872–893.

18. Chopra S., Hadsell R., LeCun Y. Learning a similarity metric discriminatively, with application to face verification. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 539–546.

19. Han J., Morag C. The influence of the sigmoid function parameters on the speed of backpropagation learning. *From Natural to Artificial Neural Computation*. Ed. by J. Mira, F. Sandoval. *Lecture Notes in Computer Science*, vol. 930. Springer, 1995, pp. 195–201.

20. Maron M.E. Automatic indexing: An experimental inquiry. *Journal of the ACM*, 1961, vol. 8 (3), pp. 404–417.

Для цитирования:

Мамикьян М.С., Бекенева Я.А. Обработка текстовых данных в медицинских диалоговых системах на примере виртуального ассистента эндокринолога // Системы анализа и обработки данных. – 2026. – № 3 (103). – С. 75–96. – DOI: 10.17212/2782-2001-2026-3-75-96.

For citation:

Mamikyan M.S., Bekeneva Ya.A. Obrabotka tekstovykh dannykh v meditsinskikh dialogovykh sistemakh na primere virtual'nogo assistenta endokrinologa [Processing text data in medical dialogue systems using the example of a virtual assistant for an endocrinologist]. *Sistemy analiza i obrabotki dannykh = Analysis and Data Processing Systems*, 2026, no. 3 (103), pp. 75–96. DOI: 10.17212/2782-2001-2026-3-75-96.