МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РАЗРАБОТКА АЛГОРИТМА ФОРМАЛИЗАЦИИ МЕДИЦИНСКИХ ТЕКСТОВ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Сидорова Сергея Александровича

Научный руководитель:
проф. кафедры информатики и
программирования, д.т.н. _____ А.С.Фалькович

подпись, дата

Зав. кафедрой:
Зав. кафедрой информатики и
программирования
к.ф.-м.н., доцент _____ М.В.Огнева

подпись, дата

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. В современном мире медицинская информация является одной из наиболее ценных категорий данных, которая используется для диагностики, лечения и исследований. В настоящее время, с увеличением объема медицинской информации и переходом к электронной системе управления здравоохранением, проблема эффективной обработки медицинских текстов становится более значимой. Медицинская терминология обладает большим количеством специфичных терминов, сокращений и контекстуальными нюансами. Разработка методов формализации помогает преодолеть сложности в интерпретации и анализе медицинских текстов, повышая точность извлечения семантической информации.

Цель магистерской работы – разработать алгоритм формализации медицинских текстов с целью улучшения доступности, анализа и использования медицинской информации.

Поставленная цель определила **следующие задачи**:

1. Провести анализ существующих работ по теме автоматизированной обработки медицинских текстов.
2. Выявить проблемы и сложности формализации медицинских текстов.
3. Определить основные компоненты системы, необходимой для формализации медицинских текстов.
4. Описать алгоритм формализации.
5. Создать приложение для формализации медицинских текстов.
6. Описать алгоритм работы приложения.
7. Провести анализ результатов формализации.

Методологические основы формализации медицинских текстов представлены в работах А. Н. Хоружой, Р. Х. Зулкарнеева, Д. В. Козлова, Н. И. Юсуповой, И. В. Москалева, Э. Г. Григоряна.

Теоретическая значимость заключается в том, что данная работа представляет собой исследование и разработку алгоритма и программы для формализации текстов инструментальных медицинских обследований. Это позволяет улучшить процессы обработки и анализа медицинских данных, что в свою очередь способствует улучшению качества и повышению эффективности медицинских исследований.

Практическая значимость заключается в том, что алгоритм и программа могут быть использованы в качестве основы для дальнейших исследований в области компьютерной обработки и анализа медицинских данных.

Структура и объём работы.

Магистерская работа состоит из введения, 7 разделов, заключения, списка использованных источников и 6 приложений. Общий объем работы – 89 страниц, из них 56 страниц – основное содержание, включая 6 рисунков и 12 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 21 наименование.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Обзор существующих исследований» посвящен анализу современных методов автоматизированной обработки медицинских текстов:

- Алгоритм автоматического выделения жалоб пациентов из историй болезни
- Автоматизированное извлечение знаний из медицинских текстов
- Методы и модели извлечения знаний из медицинских документов
- Автоматическое извлечение и расшифровка сокращений из текстов, относящихся к конкретной предметной области

- Программная система, осуществляющая case-based reasoning для диагностирования заболеваний позвоночника
- Обработка русскоязычных неструктурированных медицинских текстов и вероятностное прогнозирование групп заболеваний

Исследования, которые связаны с формализацией медицинских текстов на русском языке, в большинстве случаев не имеют открытого доступа.

Второй раздел «Основные компоненты системы для формализации медицинских текстов» посвящен основным компонентам, необходимым в создании системы для формализации медицинских текстов. К ним относятся:

1. Модуль предобработки текста:
    - Токенизация (NLTK, pymorphy2).
    - Лемматизация с учетом медицинской терминологии.
    - Фильтрация стоп-слов (с сохранением ключевых сокращений).
2. Модуль распознавания сущностей:
    - Контекстно-зависимое выделение терминов через Bio_ClinicalBERT.
    - Интеграция медицинских онтологий для разрешения неоднозначностей.
3. База знаний:
    - Хранение структурированных диагнозов в PostgreSQL.
    - Формат хранения: атрибутивная модель (тип травмы, локализация, степень тяжести).
4. Словари предметной области:
    - Автоматическое расшифровка аббревиатур)

Третий раздел «Методы и алгоритмы формализации медицинских текстов» посвящен обзору и программной реализации алгоритмов извлечения ключевой информации из текста. Выбор алгоритмов обусловлен их разным принципом работы.

Четвертый раздел «Анализ реализованных алгоритмов» посвящен сравнительному анализу результатов работы каждого из реализованных алгоритмов. Алгоритмы тестировались на корпусе текстов, которые представляют собой записи сделанные сотрудниками скорой помощи. Основным недостатком алгоритмов TextRank и RAKE является некорректная обработка коротких терминов и сокращений. Модель KeyBERT демонстрирует лучшие результаты.

Пятый раздел «Формализация медицинских диагнозов» посвящен проектированию структуры формализованной таблицы и тестированию работы полученных от каждого из алгоритмов словарей ключевых фраз с помощью классификатора, который был разработан ранее.

Установлено 8 ключевых категорий:

- Тип травмы
- Травма головы
- Травма груди
- Травма позвоночника
- Травма таза
- Травма конечностей
- Кровопотеря
- Прочее

Выбор категорий обоснован частотой встречаемости информации. Информация, которая не относится к первым семи категориям будет относится к восьмой. Такое решение принято, чтобы избежать роста таблицы в ширину.

Были оценены результаты разбиения по каждому из алгоритмов. Основные причины некорректной классификации:

- Контекстная неоднозначность
- Проблемы с сокращениями
- Ограничения токенизации

С целью повышения точности работы была сделана интеграции словаря аббревиатур вида: ключ – категория, значение – термины. Каждый из алгоритмов увеличил свою точность.

Шестой раздел «Автоматизированная формализация медицинских диагнозов» посвящен разработке комбинированного алгоритма и системы для структурирования текстов инструментальных медицинских обследований. Был проведен анализ ограничений существующих методов извлечения ключевых фраз (RAKE, TextRank, BERT) применительно к медицинским текстам, характеризующимся:

- Высокой терминологической насыщенностью;
- Контекстной зависимостью сокращений.

Для решения этих задач предложена комбинированная модель, интегрирующая:

- Глубинный семантический анализ на основе предобученной модели Bio_ClinicalBERT, адаптированной для медицинских текстов.
- Статистическую кластеризацию терминов методом K-means.

Архитектура системы реализована на микросервисной основе:

- Фронтенд: React-приложение с интерфейсом для загрузки .xlsx-файлов и визуализации результатов;
- Бэкенд: FastAPI-сервер с асинхронной обработкой запросов;
- База данных: PostgreSQL с журналированием операций.

Алгоритм формализации включает этапы:

1. Предобработку текста (нормализация регистра, удаление цифр, токенизация, лемматизация через pymorphy2).
2. Генерацию контекстно-зависимых эмбеддингов размерностью 768 признаков с помощью Bio_ClinicalBERT.
3. Адаптивную кластеризацию терминов методом K-means с валидацией векторов (исключение NaN-значений).

4. Классификацию биграммных сочетаний через косинусное расстояние до центроидов кластеров.
5. Структурирование выходных данных.

Седьмой раздел «Анализ результатов формализации» посвящен комплексной оценке эффективности разработанной системы. Для валидации модели использовались стандартные метрики классификации: Precision (точность), Recall (полнота) и F1-score (сбалансированная F-мера). Результаты демонстрируют вариативность качества обработки по тематическим категориям:

Наивысшие показатели F1-score достигнуты для категорий:

- «Грудь» (F1=0.70 при Recall=0.68)
- «Прочее» (F1=0.70 при Precision=0.63)
- «Конечности» (F1=0.68)

Стабильные результаты (F1 ≥0.63) наблюдаются для:

- «Тип травмы» (Precision=0.63, Recall=0.68)
- «Голова» (Precision=0.69)

Области для оптимизации выявлены в категориях:

- «Таз» (Recall=0.51, F1=0.63)
- «Кровопотеря и шок» (F1=0.56 — минимальный показатель)

Ключевые выводы анализа:

- Среднее значение F1-score по всем категориям составило ≥0.65, что подтверждает превосходство комбинированного подхода (Bio_ClinicalBERT + K-means) над базовыми методами (где точность ≤64%).
- Неравномерность метрик связана с терминологической спецификой: низкий Recall для категории «Таз» обусловлен редкими формулировками в исходных данных, а снижение Precision для «Кровопотери и шока» — контекстной зависимостью терминов.

- Результаты верифицировали эффективность гибридной генерации синтетических данных: модель корректно обработала 84% вариативных формулировок.

ЗАКЛЮЧЕНИЕ

В данной работе были продемонстрированы алгоритм и программа для формализации медицинских текстов.

В ходе исследования были решены поставленные задачи:

1. Проведен анализ современных подходов к автоматизированной обработке медицинских текстов, включая методы NLP (Natural Language Processing), машинного обучения и онтологического моделирования.
2. Выявлены ключевые сложности формализации медицинских текстов, такие как вариативность терминологии, наличие синонимии и аббревиатур, а также зависимость смысла терминов от контекста.
3. Определены основные компоненты системы формализации, включая модуль предобработки текста, алгоритмы семантического анализа и механизм генерации структурированных данных.
4. Разработан алгоритм формализации, основанный на комбинации методов морфологического анализа, распознавания именованных сущностей (NER) и контекстно-зависимого сопоставления терминов с эталонными онтологиями (такими как SNOMED CT, MeSH).
5. Создано приложение, реализующее предложенный алгоритм и позволяющее преобразовывать неструктурированные медицинские тексты в формализованное представление.
6. Описана архитектура приложения, включая этапы обработки текста, извлечения сущностей и формирования выходных данных в стандартизированном формате.
7. Проведен анализ результатов формализации, который подтвердил эффективность предложенного подхода в сравнении с существующими решениями.

Результаты работы демонстрируют, что применение разработанных методов позволяет повысить точность и скорость обработки медицинских текстов. Таким образом, проведенное исследование вносит вклад в развитие методов автоматизированной обработки медицинских данных, предлагая

практическое решение для их формализации, что в долгосрочной перспективе способствует повышению эффективности работы медицинских информационных систем.

Основные источники информации:

1. Хоружая А.Н., Козлов Д.В., Арзамасов К.М., Кремнева Е.И. Анализ текстов описаний кт-исследований головного мозга с признаками внутричерепных кровоизлияний с помощью алгоритма дерева решений // Современные технологии в медицине. 2022. Т. 14, № 6. С. 34 – 41.

2. Зулкарнеев Р.Х., Юсупова Н.И., Сметанина О.Н., Гаянова М.М., & Вульфин А.М. Методы и модели извлечения знаний из медицинских документов // Информатика и автоматизация. 2022. № 21(6). С. 1169-1210.

3. Москалев И.В., Кротова О.С., Хворова Л.А. Автоматизация процесса извлечения структурированных данных из неструктурированных медицинских выписок с применением технологий интеллектуального анализа текстов // Высокопроизводительные вычислительные системы и технологии 2020. Т. 4, № 1. С. 163 – 167.

4. Григорян Э.Г., Паршин М.Н. Методы NLP для предобработки текстовых данных и выделения признаков // Бизнес и общество. 2021. № 3 (31). С. 213 – 224.

5. Сидоров С.А., Фалькович А.С. Алгоритм и программа для преобразования текстовых описаний диагнозов в таблицу // Фундаментальная и прикладная медицина. Саратов, 2022. С. 141.