УДК 004.89

ПОДХОД К РЕШЕНИЮ ЗАДАЧИ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ МЕДИЦИНСКИХ ПРОТОКОЛОВ

Галай Олеся Олеговна (Национальный исследовательский университет ИТМО)
Научный руководитель – к.т.н., преподаватель Русак Алена Викторовна
(Национальный исследовательский университет ИТМО)

Введение. Первичная обработка данных медицинских протоколов является важным этапом автоматизации механизма определения симптомов с использованием машинного обучения. Её необходимость обусловлена тем, что медицинские протоколы заполняются вручную врачами, которые в свою очередь могут допускать орфографические, грамматические и синтаксические ошибки, что может привести к некорректной обработке текста. Существующие решения исправления ошибок не являются результативными из-за специфики медицинской терминологии.

Основное содержание. Для определения подхода к решению данной задачи производился анализ текущих решений первичной обработки текстовых данных в машинном обучении. Рассмотрено пять решений обработки текстовых данных. Три работы описывают обработку медицинских данных [1] [4] [5], две – обработку данных без определенной тематики [2] [3]. Данные были представлены либо на русском [1] – [3], либо на английском языках [3] – [5].
В работах, рассматривающих обработку данных без определенной тематики, был выявлен следующий набор методов предобработки: лемматизация, токенизация, формализация, векторизация, удаление стоп-слов, извлечение ключевых слов и многословных терминов, стемминнг, коррекция ошибок посредством различных библиотек.
Обзор текущих решений для первичной обработки медицинских данных показал, что все они разработаны только для конкретного направления медицины, таких как кардиология [1], неврлогия [4] и постковидный синдром [5]. Однако, их анализ позволил выделить дополнительный метод предварительной обработки данных - выделение названий болезней [1]. Его суть заключается в замене аббревиатуры, сокращений и полных названий болезней, симптомов и синдромов на полное название, написанное через нижнее подчеркивание, если оно состоит из нескольких слов.

Заключение. На основе анализа был сформулирован подход к решению задачи, который состоит в поочередном анализе использования различных методов предобработки данных и их влияния на конечный результат. Для оценки качества используемых методов будут взяты три характеристики: время обработки одного эпикриза, так как стоит задача сокращения времени получения достоверных данных из них; точность исправления ошибок и точность обучения модели, которой в качестве входных данных будут поставляться уже предварительно обработанные эпикризы.

Список использованных источников:

1. Мецкер О.Г. Методы приобретения и формализации эмпирических знаний по данным медицинских информационных систем в задачах повышения качества лечения кардиологических пациентов: дис. … канд. тех. наук: 05.13.17 / Мецкер Олег Геннадьевич. – СПб., 2018. – 135 с.
2. Методы и системы автоматического реферирования текстов: монография / Т. В. Батура, А. М. Бакиева; Ин-т систем информатики им. А. П. Ершова СО РАН. — Новосибирск: ИПЦ НГУ, 2019. — 110 с.

3. Добренко Н.В. Методы и алгоритмы интеллектуализации проектирования технических систем посредством тематической сегментации текстов: дис. … канд. тех. наук: 05.13.06 / Добренко Наталья Викторовна. – СПб., 2018. – 130 с.
4. Classification of neurologic outcomes from medical notes using natural language processing / M. Fernandes, N. Valizadeh, H. Alabsi [и др.]. — Текст: непосредственный // Expert Systems With Applications. — 2023. — № 214.
5. KTI-RNN: Recognition of Heart Failure from Clinical Notes / Li Dengao, Ma Huiting, Li Wenjing [и др.]. — Текст: непосредственный // Tsinghua Science and Technology. — 2023. — № 28. — С. 117-130.