УДК 004.912

ОБРАБОТКА МЕДИЦИНСКИХ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Функнер А.А. (Университет ИТМО), **Балабаева К.Ю.** (Университет ИТМО), **Шайкина А.А** (Университет ИТМО)
Научный руководитель – к.т.н. Ковальчук С.В.
(Университет ИТМО)

Предсказательное моделирование активно используется в областях здравоохранения и медицины. Одним из источников данных являются медицинские тексты на естественном языке, содержащиеся в электронных медицинских картах пациентов. Целью данного исследования является разработать модули обработки медицинских текстов, используя интеллектуальные методы анализа данных.

Введение. Предсказательное моделирование активно используется в областях здравоохранения и медицины. Для улучшения моделей и увеличения их точности необходимо научиться получать новые признаки моделируемых объектов. Одним из источников таких признаков являются медицинские тексты на естественном языке, содержащиеся в электронных медицинских картах (ЭМК) пациентов. Для получения признаков из таких текстов необходимо решить несколько обязательных задач: исправление опечаток, истолкование аббревиатур, обнаружение отрицаний, извлечения времени и носителя заболевания. Для работы с медицинскими текстами на английском языке существует множество инструментов, библиотек и протоколов. Некоторые из инструментов могут быть адоптированы при наличии медицинского корпуса языка. Для русского языка сейчас только разрабатывается корпус OpenCorpora, а для медницких текстов существует лишь корпус, основанный на 120 ЭМК. Помимо этого, есть ещё морфологический анализатор слов pymorphy2. Целью данного исследования является разработать модули для решения каждой из проблем обработки медицинских текстов, используя интеллектуальные методы анализа данных.

Основная часть. На данный момент разработано три модуля для исправления ошибок, обнаружения отрицаний в текстах и тематического моделирования. Модуль для исправления опечаток использует векторное представление слов (word embedding) и способен найти ближайшее слово к данному, если в нём есть опечатка. Возможные опечатки были разделены на несколько категорий: сокращение слов, пропуск или опечатка, соединенные слова. Лучшей моделью стала fastText, для которой с использованием косинусного расстояния находилось ближайшее слово. Этот подход был объединен с использованием расстояния Дамерау-Левенштейна между строками. В данном случае модуль показал точность (overall precision) 0.86 на тестовом множестве слов из ЭМК.

Модуль по поиску отрицаний позволяет классифицировать тексты на три категории: заданная болезнь упоминается, отрицается или не упоминается. Данный модуль необходим для поиска конкретных состояний пациента в тексте, и заменяет собой поиск по ключевым словам. Для обучения этого модуля необходимы размеченные тексты по заданным заболеваниям. В качестве классификатора использовался градиентный бустинг деревьев решений. Модуль был протестирован на пяти заболеваниях: инсульт (15 % отрицаний), инфаркт миокарда (7%), артериальная гипертензия (8 %), сахарный диабет (2%), стенокардия (2%). Для всех классификаторов F-мера на тестовом множестве составила от 0,81 до 0,93.

Модуль тематического моделирования работает на основе библиотеки BigARTM, разработанной на базе аддитивной регуляризации. На множестве анамнезов болезни были выделены 25 тем: назначение анализов и их результаты (литр, кровь, ТТГ, Т4, ммоль, УЗИ), показатели сердечно-сосудистой системы (ЭКГ, ЧСС, норма, мин, ритм, ФВ), рекомендации пациенту (быть, данные, для, уровень, пациент, дальнейший), стенокардия и её симптомы

(боль, выполнить, стеноз, госпитализировать, ангинозный, артерия, синдром, КАГ), артериальная гипертензия и её синдромы (АД, повышение, лечение, максимальный, ухудшение) и др. На основе выделенных тем была обучена модель TopicTiling для тематической сегментации текстов. Результаты тематической сегментации уже могут быть использованы сотрудниками медицинских учреждений для навигации по медицинской истории пациента. Также, для последующих модулей обработки медицинских текстов идентифицированная тема позволяет увеличить точность моделей, которые лежат в основе этих модулей.

Выводы. Таким образом, были разработаны три модуля для обработки медицинских текстов на естественном языке, используя интеллектуальные методы анализа данных. В будущем планируется улучшать точности вышеописанных моделей и разрабатывать новые модули для восстановления временной последовательности событий и извлечения носителя заболевания.

Функнер А. А. (автор) Подпись

Ковальчук С. В. (научный руководитель) Подпись