УДК 004.89

РАСПОЗНАВАНИЕ СИМПТОМОВ ЗАБОЛЕВАНИЙ В МЕДИЦИНСКИХ ЭПИКРИЗАХ С ИСПОЛЬЗОВАНИЕМ АНСАМБЛЯ ТРАНСФОРМЕРОВ

Курдюмов Дмитрий Александрович (ИТМО)
Научный руководитель – к.т.н., Русак Алена Викторовна (ИТМО)

Введение. Точное распознавание симптомов заболевания в медицинских документах влияет на эффективность работы систем поддержки принятия врачебных решений (СППВР), которые призваны помочь медицинскому персоналу оказать полную и направленную на конкретного человека помощь в постановке диагноза и составления индивидуальной траектории лечения. Наиболее результативные решения используют различные реализации моделей на базе архитектуры трансформер [1][2]. Однако, такие модели зависят от данных, полученных во время так называемой «адаптации предметной области», что вносит сдвиг в выделение тех или иных симптомов. Разработанный метод ансамблирования нескольких моделей может помочь смягчить индивидуальные погрешности моделей и охватить более широкий спектр ассоциаций симптомов и заболеваний. Этот подход направлен на использование разнообразия источников данных и, соответственно, повышение общей эффективности прогнозирования.

Основная часть. Извлечение симптомов в медицинских документах является задачей распознавания именованных сущностей. Чаще всего эта задача использует для разметки целевого текста IOB-нотацию, при использовании которой всем словам в документе присваивается один из трёх классов: Begin – начало именованной сущности, первое её слово; Inside – внутренняя часть именованной сущности; Outside – все остальные слова в документе, не относящиеся к именованным сущностям. Исходный датасет, используемый для оценки эффективности метода ансамблирования моделей, содержал 145 документов. Каждый из документов был представлен в формате XML, разметка именованных сущностей производилась в связанном документе в формате JSONL, где был указан xPath (идентификатор секции в xml-документе) эпикриза; индекс начала симптома и его окончание в данной секции документа. Предоставленные данные были переведены в IOB – разметку с детализацией по каждому слову в отдельных секциях эпикриза. Для создания ансамблевой модели были использованы несколько существующих свободно доступных трансформеров, на которых предварительно была выполнена задача адаптации предметной области с использованием двух различных медицинских корпусов.

Модель MedRuRobertaLarge использовалась во время создания программного обеспечения для исправления опечаток в медицинских текстах [3]. Для обучения использовалось два открытых набора данных, которые содержат в общей сложности 29 967 медицинских записей. Модель RuBioRoBERTa [4] была предварительно обучена на обучающем корпусе, состоящем из 338 тысяч статей из базы данных «КиберЛенинка»[1].

Предлагаемая ансамблевая модель направлена на предсказание симптомов из медицинского текста, полученных двумя разными моделями, путем сравнения результатов обеих моделей и выявления перекрывающихся текстовых блоков. Для достижения этой цели ансамбль использует подход сравнения расстояний позиций средних маркеров идентифицированных именованных сущностей. Далее следует пошаговое описание реализованного алгоритма работы модели ансамбля:

1) Прогнозирование симптомов: ансамблевая модель пропускает входной текст через две отдельные модели, в каждой из которых выполняется токенизация текста.
2) Распознавание симптомов: обе модели извлекают именованные сущности из текста эпикриза.

[1] https://cyberleninka.ru/

3) Расчёт позиции симптома в документе: для всех распознанных симптомов вычисляется позиция относительно начала документа как среднее между начальным и конечным индексом распознанной сущности.
4) Поиск перекрывающихся симптомов: модель ансамбля сравнивает вычисленные позиции между предсказанными именованными сущностями из двух разных моделей. Если расстояние меньше предопределенного порогового значения α (которое может зависеть от конкретного языка и средней длины слова в нём), модель ансамбля определяет, что именованные сущности перекрываются и ссылаются на одно и то же понятие.
5) Генерация консенсуса: ансамблевая модель генерирует комбинированный набор именованных сущностей на основе перекрывающихся симптомов. Этот набор представляет собой окончательный вывод ансамблевой модели. При этом полный набор токенов именованной сущности является объединением предсказанных токенов.

Выводы. Полученная ансамблевая модель сравнивалась с двумя самостоятельными моделями при помощи метрики оценки точности выделения последовательностей seqeval. F1-мера данной метрики для ансамбля моделей показала результат 0.702 на задаче определения симптомов, в то время как модель RuBioRoBERTa на этом же датасете показала результат 0.737. Полученные результаты показывают, что текущая реализация ансамблевой модели показывает сравнимую точность решения с самостоятельными моделями, однако страдает от нескольких аспектов, которые ведут к ухудшению работы:

- ручной подбор параметра максимально допустимой дистанции α между схожими именованными сущностями;
- константное значение отнесения слова к именованной сущности для каждой из моделей;
- фиксированная стратегия объединения токенов для идентичных блоков текста.

Таким образом, требуются дополнительные исследования эффективности ансамблевой модели, чтобы достоверно говорить об уровне её относительной эффективности в сравнении с самостоятельными моделями. Дальнейшее направление работы связано с созданием фреймворка для обучения приведённых выше параметров с целью создания более эффективного ансамблевого решения задачи выделения симптомов из эпикриза.

Список использованных источников:

1. Hierarchical label-wise attention transformer model for explainable ICD coding / L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett and L. Jorm. // Journal of Biomedical Informatics. — 2022. — № 133.
2. Natural Language Processing for Automated Classification of Qualitative Data From Interviews of Patients With Cancer / C. Fang, N. Markuzon, N. Patel and J. Rueda. // Value in Health. — 2022. — № 25. — С. 1995-2002.
3. RuMedSpellchecker: Correcting Spelling Errors for Natural Russian Language in Electronic Health Records Using Machine Learning Techniques / D. Pogrebnoi, A. Funkner, S. Kovalchuk. // ICCS 2023. Lecture Notes in Computer Science. — 2023. — № 10475.
4. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining / A. Yalunin, A. Nesterov, D. Umerenkov. // arxiv.org. — URL: https://arxiv.org/pdf/2204.03951.pdf (дата обращения: 16.01.2024).