Поддержка модели превентивной медицины: модуль обработки естественного языка для дистанционного взаимодействия «клиника-пациент»

А. В. Чижик[1,2], М. П. Егоров [1],
М. Ю. Якубова[1], Д. А. Погребной[1], А. С. Кривошапкина[1]

[1] Университет ИТМО, [2] Санкт-Петербургский государственный университет

chizhik@itmo.ru, egorovm@niuitmo.ru, shentorin@gmail.com,
pogrebnoy.inc@gmail.com, aitalina.kr@gmail.com

Аннотация

Модуль обработки естественного языка для дистанционного взаимодействия «клиника-пациент» является важным инструментом в поддержке модели превентивной медицины, так как позволяет улучшить качество обслуживания пациентов и повысить степень их участия в предупреждении заболеваний. В статье описывается разработанный модуль, с помощью которого можно детектировать симптомы и их отрицания, затем на этом основании выносить предварительный диагноз и маркер срочности приема. Авторами описывается общий алгоритм, особенности моделей машинного обучения, которые вкачены в общий конвейер работы модуля, приводятся метрики качества.

Ключевые слова: здравоохранение, разговорный ИИ, языковые модели, человекомашинный диалог

Библиографическая ссылка: Чижик А. В., Егоров М. П., Якубова М. Ю., Погребной Д. А., Кривошапкина А. С. Поддержка модели превентивной медицины: модуль обработки естественного языка для дистанционного взаимодействия «клиника-пациент» // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединенной научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб.: Университет ИТМО, 2024. С. 90–96. DOI: 10.17586/2541-9781-2024-7-90–96

1. Введение

Здоровье населения является одним из факторов, влияющих на полюс, к которому стремится социальное настроение, и с этой позиции оно является достаточно сложным феноменом, определяющимся через взаимодействие социальных, психологических, экономических и только затем биологических, генетических и физиологических факторов. Иными словами, во многом уровень здоровья населения зависит от мировоззренческого контекста, присутствующего в обществе и способствующего выстраиванию постоянной и доверительной коммуникации с медицинскими учреждениями. Предпосылкой для проведения этого исследования стал опыт создания мультимодального интеллектуального помощника для автоматизации процесса приема пациентов и оказания первичной медицинской помощи. Системы здравоохранения во всем мире используют автоматизацию для решения проблемы нехватки персонала, а также для эффективного управления и сортировки пациентов в больших масштабах в больницах и клиниках. Одним из видов автоматизации являются чат-боты и сопутствующие технологии, позволяющие автоматизировать процесс первичного общения с потенциальным

пациентом (включая фиксацию симптомов на стороне клиники и предоставление человеку необходимой справочной информации). Это позволяет преодолеть географические и временные барьеры между службами здравоохранения и их пользователями. Таким образом, конечной целью этих усилий является переход от неотложной помощи к профилактической, что возможно только на основе управления взаимодействием на основе данных. Важно отметить, что есть новые социологические исследования, которые показывают тенденцию: каждый пятый врач покидает профессию в течение двух лет по причине профессионального выгорания, что объясняется большим количеством рутинной деятельности (преимущественно офисного характера) [1, 2]. Чат-боты могут облегчить нагрузку на врачей, медсестер и других медицинских работников, автоматизируя задачи, которые лучше подходят для компьютера, и в результате освобождая медицинские бригады для более продуктивного выполнения своей основной работы.

Проведя серию экспериментов по созданию чат-бота с открытым доменом для общения пациента с клиникой, мы поняли, что невозможно создать полностью универсальный диалоговый агент, который любая клиника могла бы запустить в работу без дополнительных усилий. В то же время стало понятно, что клиникам нужны готовые модули, из которых можно было бы собрать нужную функциональную конфигурацию диалогового агента непосредственно на стороне клиники.

Поэтому мы решили создать библиотеку для языка Python, которая могла бы:

- обнаружить симптомы в реплике пользователя (на вход языковая модель получает короткий текст на ЕЯ, содержащий ответы со стороны пользователя на вопросы бота порядка «опишите свое самочувствие»);
- использовать эту информацию для вынесения интерпретируемого суждения о возможном диагнозе (т. е. диагноз и сопроводительную информацию о вероятностном распределении — концепция «второго мнения»);
- присвоить пациенту метку срочности (мультиклассификация пациентов для распределения потоков пациентов в клинике).

На рис. 1 показана логика разработанного модуля.

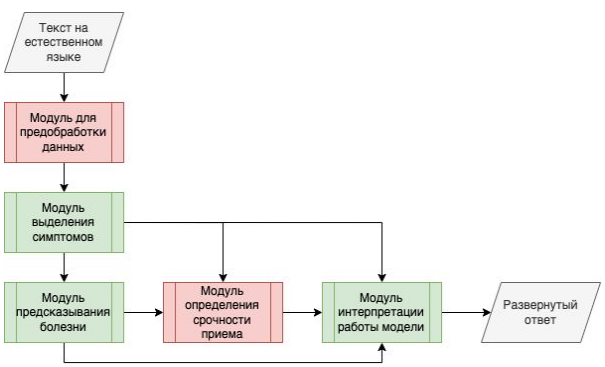

Рис. 1. Блок-схема разрабатываемого модуля

В рамках поднимаемой проблемы актуальными являются исследования, посвященные методам обработки и понимания естественного языка (NLP, natural language processing и NLU, natural language understanding). При разработке диалогового агента обычно решаются следующие классические задачи обработки естественного языка: сегментация, токенизация и лемматизация, NER и нахождение семантических связей [2]. Существуют и специфические задачи [3]: обработка последовательности из нескольких фраз, дополняющих друг друга; поиск ссылок с одной фразы на другую; обработка чередования разных типов интентов подряд; генерация уточняющих вопросов и их обработка. В нашем исследовании мы в первую очередь сосредоточились на проблеме выявления симптомов

(и их отрицания) и последующем использовании этой информации для определения диагноза [4, 5]. Следует отметить, что задача создания специализированных языковых моделей является достаточно динамично развивающейся, в частности, можно упомянуть следующие два современных исследования (относящихся к области медицины) [6, 7, 8], подход которых заключается в использовании условно закрытых данных (электронные медицинские карты пациентов, ЭМК).

2. Данные

Ролевая модель дистанционного взаимодействия между клиникой и пациентом подразумевает, что диалог строится в формате разговорного русского языка, свойственного социальным сетям (так как сам интерфейс любого диалогового агента напоминает мессенджер). Таким образом, при формировании набора текстовых данных необходимо стремиться к близости собираемых реплик к языку пациентов, а не к служебному языку медиков. Поэтому мы решили отойти от общей тенденции использования при разработке подобных модулей текстовых данных, взятых из ЭМК (анамнез и диагнозы), и собрали данные из открытых веб-источников:

- основа набора данных — 5 193 описания пациентов своих заболеваний с маркером категории болезни (источник: https://meduniver.com);
- датасет дополнен 292 заболеваниями с их описанием (источник: https://health.mail.ru/disease/adneksit/);
- также были собраны данные о симптомах из Википедии (272 симптома).

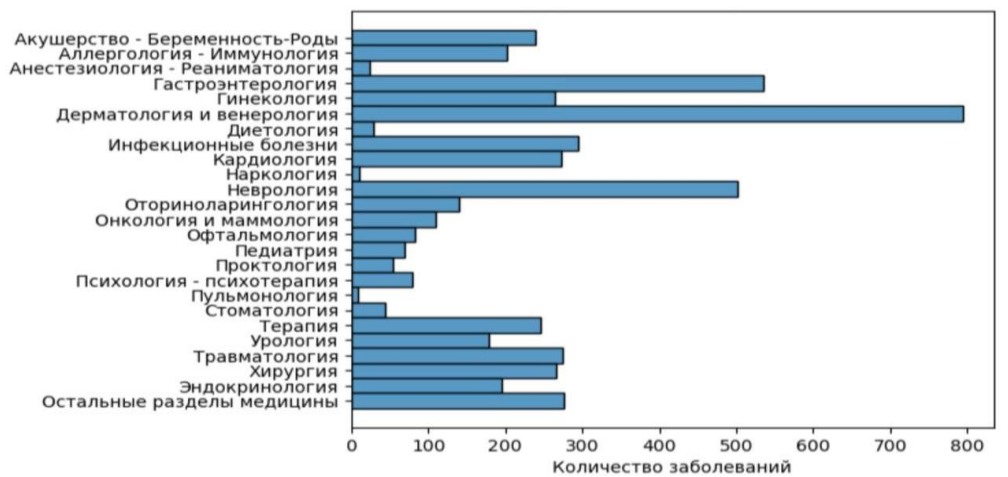

Рис. 2. Распределение категорий болезней

На рис. 2 показано распределение заболеваний по медицинским категориям. В ходе анализа данных, было выяснено, что в среднем пациент упоминает 2-3 симптома, присутствующих в его самочувствии, и тратит на описание около 66 слов.

3. Метод

На рис. 3 показана логика взаимодействия с текстовыми данными, которой мы придерживались при разработке данного модуля.

Было решено разработать систему подмодулей, что должно обеспечить возможность использовать библиотеку не только в полном функционале, но и частично, например, только для выделения симптомов.

Рис. 3. Пайплайн обработки и классификации медицинских текстов: текущая реализация

Процесс предобработки данных в рамках нашего проекта практически ничем не отличался от стандартного набора процедур, однако в отличие от классических подходов к этому этапу, мы решили сохранять некоторые стоп-слова, чтобы не потерять отрицание симптома (исходя из того, что отрицание симптома тоже является симптомом).

Далее нам потребовался список симптомов для их последующего извлечения из текстов. Хотя существуют методологии детекции ключевых слов, которые можно применить, в нашем случае они оказались не очень полезными. Поэтому был использован готовый список симптомов из открытой базы знаний «Википедия». Далее для формирования необходимой информации мы использовали фреймворк Scapy. Он может извлекать необходимые объекты из текста, используя предварительно обученную модель машинного обучения. ML-модель, доступная во фреймворке по умолчанию, не смогла справиться с большинством симптомов. Поэтому был создан некоторый набор правил-подсказок, чтобы помочь модели. Каждая такая подсказка — паттерн, написанный отдельно для каждого симптома. Как уже было отмечено выше, перед нами стояла задача детектирования отрицания симптомов. Для этой цели мы использовали пакет Python negex, который работает со всеми найденными сущностями и пытается найти отрицание для каждой из них (это реализуется за счет определения границ частей предложения и поиска в этих границах специальных слов и других признаков, полезных для задачи поиска отрицания).

После этих двух шагов у нас появился общий анамнез, сформированный на основании данных основного датасета, который включил все возможные симптомы со статусами (yes, no, no_info, confused). Диаграмма этого процесса представлена на рис. 4.

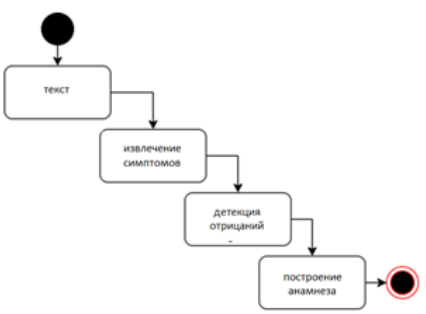

Рис. 4. Логика формирования набора симптомов со статусами

В моменте взаимодействия с репликой пользователя появляется задача мультиклассовой классификации, которая в нашем модуле решается с помощью модели логистической регрессии.

Отметим, что логистическая регрессия — это статистическая модель, которая используется для прогнозирования вероятности возникновения некоторого события, в данном случае для диагностики конкретного заболевания. Модель применяется к подготовленным данным и подразумевает создание матрицы признаков, в которой каждый столбец представляет определенный симптом, а каждая строка представляет конкретный случай пациента. Значения в ячейках матрицы указывают на наличие (1), отсутствие (0) или отрицание (-1) симптома. Используя подготовленную матрицу признаков и соответствующие метки классов (заболеваний), модель логистической регрессии обучается. В процессе обучения модель определяет оптимальные веса для каждого признака (симптома), которые позволяют наиболее точно классифицировать заболевания.

Преимуществом логистической регрессии является интерпретируемость результатов. Веса, присвоенные каждому симптому, отражают важность этого симптома в определении заболевания. Таким образом, врачи и другие медицинские работники могут анализировать эти веса и понимать логику результата модели. Более того, логистическая регрессия учитывает наличие, отсутствие и отрицание симптомов, что делает результаты еще более точными и надежными. Из вышеизложенного ясно, что нам становится легко получить удобочитаемую интерпретацию диагноза пациента. Текущая точность модели составляет 86%.

Можно обозначить конечной целью нашего модуля определение срочности приема/госпитализации пациента. Система здравоохранения в России предполагает 3 формы помощи: экстренную, срочную и плановую. Поэтому, чтобы разметить данные по срочности, мы собрали симптомы из различных открытых источников по категориям экстренных, срочных и плановых приемов. Для тестирования точности наших моделей машинного обучения мы сосредоточились на заболеваниях из категории «кардиология». Размеченные на три класса срочности симптомы были верифицированы на предмет применимости к задаче кардиологами Национального медицинского исследовательского центра им. В. А. Алмазова. Отметим, при отнесении случая к первым двум категориям (экстренный и срочный прием) пациенту требуется госпитализация, поэтому было принято решение маркировать данные бинарно: госпитализация требуется или не требуется.

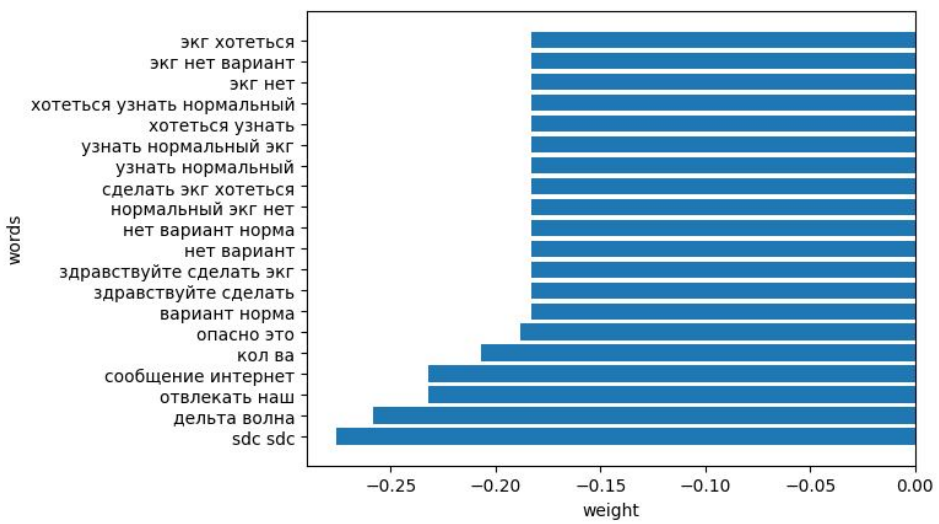

Рис. 5. Содержания класса «плановый прием»

Из предобработанных симптомов были составлены биграммы и триграммы. Дальнейшая логика была такова: если в предобработанном тексте пациента присутствовал хотя бы один из этих 2-3-граммов, то кейс маркировался как срочный. В результате мы получили датасет, содержащий 41 срочный случай и 231 плановый. Базовой идеей было использование модели логистической регрессии и tf-idf векторизатора. Однако модель переобучилась из-за несбалансированности классов и присутствию в «несрочном» классе большого количества шума (класс содержал просьбы пациентов интерпретировать результаты анализов) — рис. 5.

Поэтому следующим шагом стало обучение нейросети на Self Attention. Для каждого слова были взяты word2vec-эмбеддинги, затем Self Attention был использован для анализа контекста каждого слова. В итоге получены следующие метрики качества модели маркировки срочности: Accuracy = 0.93 и F1 = 0.96.

4. Заключение

В настоящее время коллектив авторов работает над улучшением значений метрик качества и планирует измерять качество модуля за счет привлечения медицинских экспертов для тестирования. Кроме того, наборы данных планируется дополнить новыми случаями. На наш взгляд, текущие тесты показывают, что модуль применим на практике. Наборы данных и библиотека находятся в свободном доступе на github (https://github.com/NIRMA-PATIENT-INTAKE).

Исследование проведено в рамках НИР Университета ИТМО № 622275 «Разработка модуля для предсказания предварительного диагноза: поддержание логистики потоков пациентов и концепции второго мнения при взаимодействии с пациентом через диалоговые системы».

Литература

[1] Sinsky C. A., Brown R. L., Stillman M. J., Linzer M. COVID-Related Stress and Work Intentions in a Sample of US Health Care Workers // Mayo Clinic Proceedings: Innovations, Quality & Outcomes. 2021. Vol. 5 (6). P. 1165–1173.

[2] Sinsky C., Colligan L., Li L., Prgomet M., Reynolds S., Goeders L., Westbrook J., Tutty M., Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties // Ann Intern Med. 2021. Vol. 165 (11). P. 753–760. DOI: 10.7326/M16-0961.

[3] Lalwani T. et al. Implementation of a Chatbot System using AI and NLP // International Journal of Innovative Research in Computer Science & Technology (IJIRCST). 2018. Vol. 6 (3). P. 26–30. DOI: 10.2139/ssrn.3531782.

[4] Jurafsky D., Martin J. H. Title Speech and Language Processing. 2nd edition. Prentice Hall, 2008.

[5] Freedman M. S., Gray T. A. Vascular headache: a presenting symptom of multiple sclerosis // Canadian journal of neurological sciences. 1989. Vol. 16 (1). P. 63–66.

[6] Chizhik A., Egorov M. Multimodal Intelligent Assistants for Automating the Patient Intake Process and Primary Care // 15th International Conference on Theory and Practice of Electronic Governance. — 2022. — P. 573–575.

[7] Legnar M. et al. Natural Language Processing in Diagnostic Texts from Nephropathology // Diagnostics. — 2022. — Vol. 12 (7). — Art. 1726. DOI: 10.3390/diagnostics12071726.

[8] Zhou B. et al. Natural language processing for smart healthcare // IEEE Reviews in Biomedical Engineering. — 2022. DOI: 10.48550/arXiv.2110.15803.

Support for the Preventive Medicine Model: Natural Language Processing Module for Remote Clinic-Patient Interaction

Anna V. Chizhik [1,2], Michil P. Egorov [1], Maria Yu. Yakubova [1], Dmitrii A. Pogrebnoi [1], Aitalina S. Krivoshapkina [1]

[1] ITMO University, [2] St.Petersburg State University

The natural language processing module for remote clinic-patient interaction is an important tool in supporting the model of preventive medicine, as it allows you to improve the quality of patient care and increase their degree of participation in disease prevention. The article describes the developed module, which can be used to detect symptoms and their denials, then, on this basis, make a preliminary diagnosis and a marker of the urgency of admission. The authors describe the general algorithm, the features of machine learning models that are injected into the general pipeline of the module, and provide quality metrics.

Keywords: health service, conversational AI, language models, human machine dialogue

Reference for citation: Chizhik A. V., Egorov M. P., Yakubova M. Yu., Pogrebnoi D. A., Krivoshapkina A. S. Support for the Preventive Medicine Model: Natural Language Processing Module for Remote Clinic-Patient Interaction // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). - St. Petersburg: ITMO University, 2024. P. 90–96. DOI: 10.17586/2541-9781-2024-7-90–96

Reference

[1] Sinsky C. A., Brown R. L., Stillman M. J., Linzer M. COVID-Related Stress and Work Intentions in a Sample of US Health Care Workers // Mayo Clinic Proceedings: Innovations, Quality & Outcomes. 2021. Vol. 5 (6). P. 1165–1173.

[2] Sinsky C., Colligan L., Li L., Prgomet M., Reynolds S., Goeders L., Westbrook J., Tutty M., Blike G. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties // Ann Intern Med. 2021. Vol. 165 (11). P. 753–760. DOI: 10.7326/M16-0961.

[3] Lalwani T. et al. Implementation of a Chatbot System using AI and NLP // International Journal of Innovative Research in Computer Science & Technology (IJIRCST). 2018. Vol. 6 (3). P. 26–30. DOI: 10.2139/ssrn.3531782.

[4] Jurafsky D., Martin J. H. Title Speech and Language Processing. 2nd edition. Prentice Hall, 2008.

[5] Freedman M. S., Gray T. A. Vascular headache: a presenting symptom of multiple sclerosis // Canadian journal of neurological sciences. 1989. Vol. 16 (1). P. 63–66.

[6] Chizhik A., Egorov M. Multimodal Intelligent Assistants for Automating the Patient Intake Process and Primary Care // 15th International Conference on Theory and Practice of Electronic Governance. — 2022. — P. 573–575.

[7] Legnar M. et al. Natural Language Processing in Diagnostic Texts from Nephropathology // Diagnostics. — 2022. — Vol. 12 (7). — Art. 1726. DOI: 10.3390/diagnostics12071726.

[8] Zhou B. et al. Natural language processing for smart healthcare // IEEE Reviews in Biomedical Engineering. — 2022. DOI: 10.48550/arXiv.2110.15803.