Обработка текстовой медицинской информации: метод сбора и маркировки симптомов заболеваний

А.В. Чижик[1], С.А. Мельникова[2]

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

a.chizhik@spbu.ru, melnikova-rostovskaya@yandex.ru

Аннотация

Данная статья посвящена описанию проблем при сборе обучающего набора данных для создания диалогового агента для медицинской сферы. Авторами описывается собственный эксперимент, заключающийся в разработке стратегии выкачивания данных из сети интернет и последующего анализа слабоструктурированных текстов с целью выделения фактов и именованных сущностей. Итогом проведенного исследования является опубликованный датасет.

Ключевые слова: датасет, NER, анализ текста

Библиографическая ссылка: Чижик А.В., Мельникова С.А. Обработка текстовой медицинской информации: метод сбора и маркировки симптомов заболеваний // Компьютерная лингвистика и вычислительные онтологии. Выпуск 6 (Труды XXV Международной объединенной научной конференции «Интернет и современное общество», IMS-2022, Санкт-Петербург, 23 – 24 июня 2022 г. Сборник научных статей). — СПб.: Университет ИТМО, 2022. С. 12-19. DOI: 10.17586/2541-9781-2022-6-12-19

Введение

Практически все существующие задачи машинного обучения базируются на необходимости наличия обучающего набора данных, в особенности это касается задач классификации и выстраивания человеко-машинного диалога. Датасет – это обработанные и структурированные данные в табличном виде. Иными словами, это размеченные данные, на основе которых и происходит машинное обучение. Например, если перед нами стоит задача интеграции взаимодействия голосовыми командами с интерфейсом, то для начала своей модели мы должны предоставить набор данных с транскрипциями живой речи. Чем больше образцов речи нам удастся собрать в наш датасет, тем лучше модель будет отрабатывать n+1 случай.

На данный момент гипотеза возможности внедрения ИИ в рутинный конвейер взаимодействия «клиника-пациент» выглядит достаточно привлекательной. Однако, есть ряд ограничений: во-первых, все еще нет четкой позиции медиков и юристов на тему легитимности включения ИИ в рутинные процессы на первом этапе взаимодействия пациента и клиники; во-вторых, практически всегда реализация общения «робота» с живым человеком строится на интерфейсах с использованием правил (если…,то…), а не на реализации человеко-машинного разговора в формате естественного диалога. Последняя проблема является ключевой для нас. Отметим, что для того, чтобы построить коммуникацию на естественном языке и при этом преследовать цель сбора полезных фактов, а также не разочаровывать бессмысленностью ответов, необходима разработка сценариев поведения чат-бота, но, главное, требуется сбор датасетов с примерами как реплик человека (для детекции классов намерения пользователя), так и с коллекцией симптомов (ведь основная задача первичного взаимодействия пациента с клиникой почти

всегда сводится к тому, что человек хочет записаться на прием к врачу, а регистратура нуждается в информации на тему его состояния).

Также стоит упомянуть перспективную задачу анализа поведения пациентов вне клиники с точки зрения рефлексии на тему своего самочувствия и поставленных диагнозов. Такая проблема также упирается в сбор медицинских текстовых данных и в последующем нахождения в них закономерностей.

Базовым подходом к сбору медицинских данных является обращение к медицинским выпискам. В целом это логично, так как на данный момент любой приход к врачу фиксируется в цифровой форме, что означает наличие после каждого наблюдения пациента большого количества данных: информация о физическом состоянии, симптомы, описание жалоб и т.д. Однако такого рода информация ограничена в распространении, использование ее в чат-ботах является сомнительным кейсом (так как нет согласия автора текстов, т.е. пациента). В связи с этим возникает гипотеза, что стоит обратиться к анализу большого количества данных, размещенных в сети Интернет.

Разделы сайтов клиник типа «вопрос-ответ», а также взаимодействия между людьми в социальных сетях с целью обмена публичной информацией представляют собой важный новый источник данных.

Стоит отметить, что количество данных, содержащихся в открытых источниках (новые и социальные медиа), можно отнести к категории Big Data [1].

Если обратиться к рефлексии на тему полезности такого рода слабо структурированных данных, то их можно использовать для понимания воздействия лекарств, болезней и методов лечения на пациентов за пределами контролируемых клинических условий, а также, чтобы понять паттерны поведения индивидов, связанные со здоровьем.

Для нас же было интересно попытаться выделить из этого массива данные, которые бы можно было использовать для разметки симптомов на классы (определение диагноза).

Особенность интернета как источника таких данных заключается в том, что перед нами оказываются тексты на естественном языке с очень разнообразной структурой. Это их существенно отличает от электронных медицинских записей, где содержится текст на естественном языке, однако он более формален, что дает возможность быстрее извлечь и структурировать релевантную информацию.

1. Постановка задачи

Перед нами стояла задача «чтения» неструктурированных текстов с целью сбора определенной медицинской информации и формирования на ее основе датасета, пригодного для использования в алгоритмах машинного обучения.

Датасет отличается от простого сбора медицинских данных тем, что он наделён особыми свойствами — унификацией и структурированностью данных; отсутствием грубых неточностей; наличием дополнительной информации. Каждый датасет уникален не только образцами данных, которые в него входят, но также способом их классификации и подходами к разметке.

Для того чтобы создать датасет, которым можно воспользоваться в дальнейших исследованиях и применить в прикладных разработках, было необходимо сформулировать клиническую и/или практическую задачу в области медицины, которая (потенциально) подвержена автоматизации с помощью интеллектуальных систем.

В нашем случае прикладной целью создания датасета явилось желание создать сервис для социологических опросов на медицинскую тему. Веб-сервис предполагается реализовать в формате чат-бота с открытым доменом для симуляции человеко-машинного диалога (симуляция медицинского сервиса). Основной конечной целью такой прикладной разработки мы видим изучение пользовательского опыта и фиксацию пути решения стандартной проблемы при обращении в автоматизированные центры скорой помощи

и регистратуры (так как мы предполагаем, что положительное восприятие автоматизации происходит за счет использования естественного языка со стороны бота вместо реализации подобных приложений на базе правил). Таким образом, датасет с размеченными классами симптомов будет полезным для решения технических задач, в том числе распознавания намерения пользователя. Также мы хотим протестировать с помощью респондентов точность классификации, которая возможна при условии наличия «чистого» датасета.

Стоит отметить, что для таких задач существуют специализированные сервисы, например Amazon Comprehend Medical [2]. Он позволяет извлекать значимую информацию (жалобы, диагноз, назначенные препараты и их дозировку, результаты исследований и т.п) из неструктурированных медицинских записей. В конце 2019 г. JAMIA [3] опубликовала исследование, которое показало, что данные, полученные из неструктурированных ЭМК, являются более точным источником информации для прогнозирования ИБС, чем структурированные данные. Это подкрепляет нашу перспективность извлечения признаков из неструктурированных открытых медицинских данных для задач сбора больших наборов данных для машинного обучения.

Для решения нашей задачи она была разбита на три подзадачи:

- выбор источников данных и сбор чернового датасета (неструктурированные тексты и техническая информация);
- извлечение названий болезней (классы);
- выделение симптомов.

2. Этап отбора исходных данных

Медицинские данные могут накапливаться как при фиксации рутинного диагностического процесса в медицинском учреждении, так и с использованием неструктурированных данных из социальных и новых медиа. Люди часто обеспокоены состоянием своего здоровья и рядом медицинских проблем, особенно когда речь идет о сложных или хронических заболеваниях. Пациенты часто желают иметь легкий доступ к информации о заболеваниях и симптомах, чтобы понять свое состояние и облегчить самоконтроль заболеваний, не полагаясь полностью на взаимодействие с врачом [4]. К примеру, пациенты с хроническими заболеваниями используют социальные сети для получения эмоциональной и практической поддержки [5]. К тому же медицинские работники часто делятся своим опытом в социальных сетях, в том числе исходя из маркетинговых целей. Существует статистика, согласно которой 42% интернет-пользователей используют социальные сети для получения медицинской информации. 29% ищут информацию о здоровье через платформы социальных сетей, чтобы наблюдать за опытом других пациентов с их заболеваниями [6] (это значит, что в одном сообщении появляются как диагнозы, так и симптомы). Разговоры пользователей на темы, связанные со здоровьем, которые содержатся в Twitter и Facebook, были проанализированы рядом ученых. В частности, есть исследование, описывающее анализ текстовых данных с целью выделения кластеров симптомов рака молочной железы [7]. Также текстовые данные изучали в контексте анализа поведения курящих людей [8, 9].

Еще одна научная группа [10] извлекла данные из онлайн-сообществ, посвященных вопросам здравоохранения, ученые использовали кластеризацию текстовых данных для изучения потребностей и интересов пациентов. Были извлечены медицинские термины, в том числе связанные с состояниями, симптомами, лечением, эффективностью и побочными эффектами. Их результаты показывают, что существовали значительные различия в темах, которые присутствовали на различных платформах для обсуждения болезней, если сравнивать их с вектором популярного контента в научной литературе.

Латентное распределение Дирихле (LDA) было использовано для кластеризации обсуждений диагнозов в Facebook* [11]. Сгруппированные тематические кластера были проанализированы с точки зрения полярности настроений. Таким образом, становится очевидно, что данные, собранные из открытых источников, обычно имеют широкую вариативность в изложении информации и позволяют создать наиболее репрезентативный датасет. Ключевое значение имеет баланс классов: в какой пропорции распределены образцы в датасете. Проблема работы с несбалансированными наборами данных заключается в том, что большинство методов машинного обучения игнорирует или имеет низкую производительность при анализе малого класса. В то время как с практической точки зрения производительность в этом классе является наиболее важной задачей.

В задаче классификации данные называются несбалансированными, если в обучающей выборке доли объектов разных классов существенно различаются [12]. Проблема с дисбалансом чаще всего возникает, когда какой-то из классов соответствует очень редко наблюдаемым или диагностируемым явлениям (например, редкая болезнь). Для решения задачи уравновешивания классов обычно применяют перебалансировку данных: недо- или пересэмплирование (SMOTE). SMOTE заключается в идее увеличения малого класса за счёт представителей выпуклых комбинаций пар. Самый простой подход включает в себя дублирование примеров в миноритарном классе, хотя эти примеры не добавляют в модель никакой новой информации. Поэтому новые примеры могут быть синтезированы из существующих примеров. Этот тип увеличения данных для малого класса называется методом пересэмплирования. Вторая идея перебалансировки [13] заключается в том, что мы можем сделать классы сбалансированными за счет замены большого класса подвыборкой по мощности равной малому классу, этот прием называется недосэмлированием. Подводя практический итог, простейшая стратегия недосэмплирования – взять случайную подвыборку, простейшая стратегия пересэмплирования – продублировать объекты малого класса. У пересэмплирования качество, как правило, выше, т.к. используются все данные, однако недосэмплирование позволяет учить модель на маленькой выборке.

3. Обзор возможностей извлечения качественных характеристик заболевания и описание полученных результатов

На первом этапе мы сформировали набор из 120 заболеваний, которые взяли на официальных сайтах клиник. Эта информация была необходима для того, чтобы затем собирать описания симптомов этих заболеваний. Таким образом, мы выявили метки классов и получили по одному образцу статьи, описывающей заболевание. Средняя длина такой статьи 4000 знаков. Далее было принято решение сфокусироваться при сборе данных на 5 заболеваниях: «аппендицит», «холицестит», «эзофагит», «энтерит», «язва». В итоге был собран корпус из 13 624 текстовых сообщений на тему этих заболеваний (разделы сайтов медицинских учреждений, дискуссии в социальных сетях, посвященные конкретным заболеваниям). Дальнейшую задачу можно сформулировать как извлечение качественных характеристик заболевания (указывающих на симптоматику) из текста.

Для анализа медицинских текстовых записей используются специализированные методы и инструменты распознавания именованных сущностей (NER). Эти методы обработки человеческого языка позволяют находить в образцах текста на естественном языке опредмеченные категории слов и словосочетаний. Часто (как и в нашем случае) речь идет о работе с метками классов, количество которых является заранее известным. Классами могут являться наименования заболеваний, факты госпитализации пациента

*Минюст России внес американскую корпорацию Meta (является головной компанией Facebook, который также запрещен в России) в реестр экстремистских организаций.

(есть или нет), различные количественные и качественные параметры, признаки и случаи и т.д.

Большинство NER-классификаторов базируются на алгоритме CRF (Conditional random field), который относится к классу скрытых марковских моделей [14].

Для создания моделей NER-классификации обычно используется библиотека с открытым исходным кодом для обработки естественного языка – SpaCy. Она написана на языке программирования Python, выполняет токенизацию, разметку части речи (PoS) и разбор зависимостей. Библиотека опубликована под лицензией MIT. SpaCy предлагает 18 меток (tags), которыми отмечаются именованные сущности, а также простой способ дообучить свою собственную модель. В работе [15] оценивались 10 различных готовых систем извлечения признаков на точность и скорость извлечения. В итоге SpaCy показал лучшую скорость извлечения, поддерживая сопоставимую точность от 85% до 90%. Модели SpaCy представляют собой сверточные нейросети (CNN), которые позволяют делать прогноз, основанный на представленных во время обучения примерах.

В 2020 г. университет Стенфорда выпустил свою библиотеку Stanza [16]. Модули библиотеки построены на основе библиотеки PyTorch. Это набор инструментов, которые можно использовать для создания конвейеров нейронных сетей для анализа текста. Библиотека поддерживает такие функции как токенизация, расширение токена до нескольких слов, лемматизация, части речи (POS), тегирование морфологических признаков, анализ зависимостей, распознавание именованных сущностей и анализ настроений. Она использует универсальные зависимости для предоставления согласованных аннотаций грамматики на более чем 60 языках. Таким образом, по сути, библиотека покрывает минус SpaCy (отсутствие языковых моделей для некоторых языков, в том числе поддержка русского языка стала возможна только в 2021 г. и пока годится не для всех задач) и реализует мультиязычность. Отметим, что Stanza предоставляет готовые функции, которые поддерживают синтаксический анализ и распознавание именованных объектов в текстах клинических выписок (что оказалось полезно для нашей задачи).

Нами также была предпринята попытка анализа текстов с помощью проекта Natasha. Это один из главных NLP-проектов для русского языка. Он имеет долгую историю, и начинался с rule-based решений, сейчас же библиотека решает основные задачи NLP для русского языка современными методами: токенизацию, сегментация предложения, лемматизация, нормализация фразы, синтаксический разбор, NER-тегирование, извлечение фактов. К минусам можно отнести нестабильный результат работы (в зависимости от сложности и специфичности текстов, подаваемых на вход).

В итоге мы остановились на использовании библиотеки Stanza.

Для оценки качества задач NER, как правило, используются метрики: Precision (точность), Recall (полнота), F1 (среднее гармоническое точности и полноты). При этом для улучшения детекции симптомов мы проанализировали исходный датасет с использованием метрик релевантности, читаемости и спамности (которые заимствовали из маркетинговых исследований). После чего удалили тексты не соответствующие критериям читаемости (иными словами, проверили тексты на предмет присутствия бессодержательной демагогии), чтобы получаемые данные были более «читаемы» и нашим алгоритмом извлечения симптомов.

В итоге обученная нами модель имела следующие метрики качества: Precision=91,1%, Recall=87,3%, F1 = 89,2%.

Заключение

Обученная модель извлечения признаков из медицинских текстов методами NLP показала достаточную точность при обработке неструктурированных текстовых данных. Собранные данные будут тестироваться внутри чат-бота (ручными методами),

а датасет планируется продолжить пополнять. Итоговый датасет выложен в открытый доступ на GitHub[1].

Литература

[1] Ward J. S., Barker A. Undefined by data: a survey of big data definitions // arXiv preprint arXiv:1309.5821. 2013.

[2] Bhatia P. et al. Comprehend medical: a named entity recognition and relationship extraction web service // 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE, 2019. P. 1844-1851.

[3] Hernandez-Boussard T. et al. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies // Journal of the American Medical Informatics Association. 2019. Vol. 26. № 11. P. 1189-1194.

[4] Denecke K., Brooks E. Web science in medicine and healthcare // Methods of Information in Medicine. 2013. Vol. 52. № 2. P. 148-151.

[5] Patel R. et al. Social media use in chronic disease: a systematic review and novel taxonomy // The American journal of medicine. 2015. Vol. 128. № 12. P. 1335-1350.

[6] Referral M.D. 30 Facts & Stats on Social Media and Healthcare. 2017. URL: https://getreferralmd.com/2017/01/30-facts-statistics-on-social-media-and-healthcare (дата обращения: 09.05.2022).

[7] Marshall S. A. et al. Symptom clusters in women with breast cancer: an analysis of data from social media and a research study // Quality of Life Research. 2016. Vol. 25. № 3. P. 547-557.

[8] Myslín M. et al. Using twitter to examine smoking behavior and perceptions of emerging tobacco products // Journal of medical Internet research. 2013. Vol. 15. № 8. P. 2534.

[9] Pandrekar S. et al. Social media based analysis of opioid epidemic using Reddit // AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2018. Vol. 2018. P. 867.

[10] Lu Y. et al. Health-related hot topic detection in online communities using text clustering // Plos one. 2013. Vol. 8. № 2. P. 56221.

[11] Nzali M. D. T. et al. What patients can tell us: topic analysis for social media on breast cancer // JMIR medical informatics. 2017. Vol. 5. № 3. P. 7779.

[12] Gardner M. et al. Evaluating Models' Local Decision Boundaries via Contrast Sets // arXiv preprint arXiv:2004.02709. 2020.

[13] Le Bras R. et al. Adversarial filters of dataset biases // International Conference on Machine Learning. PMLR, 2020. P. 1078-1088.

[14] Rabiner L., Juang B. An introduction to hidden markov models // Ieee assp magazine Citado na. Vol. 3 (1): 4–16, 1986. P. 31.

[15] Choi J. D., Tetreault J., Stent A. It depends: Dependency parser comparison using a web-based evaluation tool // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015. С. 387-396.

[16] Qi P. et al. Stanza: A Python natural language processing toolkit for many human languages // arXiv preprint arXiv:2003.07082. 2020.

[1] https://github.com/Frantsuzova/SDH/tree/main/data

Mining of Textual Health Information: Method for Collecting and Labeling Symptoms of Diseases

A.V. Chizhik[1], S.A. Melnikova[2]

[1] Saint Petersburg State University, [2] ITMO University

Resources hosted on the Internet constitute a rich data source for natural language processing tasks such as named entity recognition, relation extraction, and sentiment analysis. In particular, such platforms about health provide a different insight into patient's experiences with diseases. It becomes possible to collect disease symptoms and compile a dataset that can serve as a basis for telemedicine applications. This paper aimed to report a study of entities related to chronic diseases and their relation in user-generated text posts.

Keywords: telemedicine, mobile health, dataset, data mining

Reference for citation: Chizhik A.V., Melnikova S.A. Mining of Textual Health Information: method for collecting and labeling symptoms of diseases // Computer Linguistics and Computing Ontologies. Vol. 6 (Proceedings of the XXV International Joint Scientific Conference «Internet and Modern Society», IMS-2022, St. Petersburg, June 23-24, 2022). - St. Petersburg: ITMO University, 2022. P. 12-19. DOI: 10.17586/2541-9781-2022-6-12-19

Reference

[1] Ward J. S., Barker A. Undefined by data: a survey of big data definitions // arXiv preprint arXiv:1309.5821. 2013.
[2] Bhatia P. et al. Comprehend medical: a named entity recognition and relationship extraction web service // 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE, 2019. P. 1844-1851.
[3] Hernandez-Boussard T. et al. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies // Journal of the American Medical Informatics Association. 2019. Vol. 26. № 11. P. 1189-1194.
[4] Denecke K., Brooks E. Web science in medicine and healthcare // Methods of Information in Medicine. 2013. Vol. 52. № 2. P. 148-151.
[5] Patel R. et al. Social media use in chronic disease: a systematic review and novel taxonomy // The American journal of medicine. 2015. Vol. 128. № 12. P. 1335-1350.
[6] Referral M.D. 30 Facts & Stats on Social Media and Healthcare. 2017. URL: https://getreferralmd.com/2017/01/30-facts-statistics-on-social-media-and-healthcare (data obrashcheniya: 09.05.2022).
[7] Marshall S. A. et al. Symptom clusters in women with breast cancer: an analysis of data from social media and a research study // Quality of Life Research. 2016. Vol. 25. № 3. P. 547-557.
[8] Myslín M. et al. Using twitter to examine smoking behavior and perceptions of emerging tobacco products // Journal of medical Internet research. 2013. Vol. 15. № 8. P. 2534.
[9] Pandrekar S. et al. Social media based analysis of opioid epidemic using Reddit // AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2018. Vol. 2018. P. 867.
[10] Lu Y. et al. Health-related hot topic detection in online communities using text clustering // Plos one. 2013. Vol. 8. № 2. P. 56221.
[11] Nzali M. D. T. et al. What patients can tell us: topic analysis for social media on breast cancer // JMIR medical informatics. 2017. Vol. 5. № 3. P. 7779.
[12] Gardner M. et al. Evaluating Models' Local Decision Boundaries via Contrast Sets // arXiv preprint arXiv:2004.02709. 2020.

[13]Le Bras R. et al. Adversarial filters of dataset biases // International Conference on Machine Learning. PMLR, 2020. P. 1078-1088.
[14]Rabiner L., Juang B. An introduction to hidden markov models // Ieee assp magazine Citado na. V. 3 (1): 4–16, 1986. P. 31.
[15]Choi J. D., Tetreault J., Stent A. It depends: Dependency parser comparison using a web-based evaluation tool // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015. S. 387-396.
[16]Qi P. et al. Stanza: A Python natural language processing toolkit for many human languages // arXiv preprint arXiv:2003.07082. 2020.