Опыт создания языкового модуля для предсказания предварительного диагноза пациента при взаимодействии с ним с использованием диалогового агента

А.В. Чижик, М.П. Егоров

***Аннотация*— В последние годы диалоговые агенты на базе искусственного интеллекта рассматриваются как перспективный метод первичного взаимодействия с пациентом при его обращении в клинику. Действительно, это помогает разгрузить регистратуру, а также оптимизировать потоки пациентов внутри медицинского учреждения. Стоит отметить, что клиентский опыт, возникший у индивидов за счет активного использования онлайн-среды для решения повседневных задач, мотивирует реализовывать диалоговые агенты таким образом, чтобы взаимодействие было быстрым и удобным. Именно за счет этого создается мотивация к дальнейшему использованию этого канала коммуникации. Однако чат-боты на русском языке на данный момент плохо справляются с интерпретацией полученных со стороны пользователя текстовых данных. Так, сложной задачей является выделение симптомов из вводимой пользователем строки, а в связи с этим существует дальнейшая проблема классификации текстов. Статья посвящена описанию процесса проектирования и разработки языковой модели для предсказания предварительного диагноза пациента при взаимодействии с использованием диалогового агента. В качестве данных для обучения моделей и валидации работы алгоритмов были выбраны априорные знания о болезнях и их симптомах, слабоструктурированные данные с форумов общего обсуждения заболеваний, а также сгенерированный набор текстов с использованием ChatGPT. В статье описывается общая идея создаваемой библиотеки, раскрывается тема классификации заболеваний, а также анализируются метрики качества разработанных моделей.**

***Ключевые слова*— классификация текстов, логистическая регрессия, языковая модель, диалоговые агенты, eHealth.**

I. Введение

Системы здравоохранения по всему миру проходят серьезные изменения благодаря развитию цифровых технологий. Врачебная практика стремительно внедряет новые медицинские технологии, которые используют искусственный интеллект. Эти инновации включают в себя различные приложения и программы, которые помогают врачам и медицинскому персоналу в диагностике, лечении и управлении потоками пациентов. Благодаря искусственному интеллекту, системы здравоохранения становятся более эффективными и персонализированными в своей работе, что позволяет улучшить качество медицинской помощи и перейти к концепции превентивной медицины. Эти усилия актуализировали работу над одним из перспективных направлений – проектирование и разработка механизмов, способных анализировать состояние пациента в моменте первого коммуникативного взаимодействия с клиникой. Подобные разработки могут помочь снизить рутинную нагрузку на медицинских работников, а также являются одним из логичных подходов к предоставлению врачам концепции «второго мнения» [1]. Развитие ИКТ (информационно-коммуникационных технологий) позволяет проводить часть коммуникативных взаимодействий с пациентами в дистанционном формате, а также разгружать колл-центры и регистратуры, используя для сбора необходимой для приема информации диалоговые интерфейсы [2, 3]. Именно этот факт является основой гипотезы, что появляется достаточно много данных со стороны пациента, которые на пути к врачу через каналы связи, могут быть обработаны алгоритмами машинного обучения.

Сам концепт телемедицины (использование автоматизации и искусственного интеллекта для взаимодействия «клиника-пациент») не является новым. Однако пандемия COVID-19 стала фактором, который стимулировал быстрое развитие технологий телемедицины. Стоит отметить, что применение телемедицинских технологий в ряде случаев способствует более справедливому распределению специализированной медицинской помощи [4].

Существует множество факторов, влияющих на успешное внедрение и эффективность телемедицины как базиса eHealth. Поддержание высокого уровня оказываемой медицинской помощи в условиях «дистанционной медицины» и получение удовлетворительных исходов лечения зависит от уровня осведомленности пациентов и медперсонала о легитимном контуре функционирования телемедицины на уровне законодательной базы (правовая база в

Статья получена 01 августа 2023.
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».
А.В. Чижик, Центр технологий электронного правительства Института дизайна и урбанистики Университета ИТМО (afrancuzova@mail.ru)
М.П. Егоров, Центр технологий электронного правительства Института дизайна и урбанистики Университета ИТМО (egorovm@niuitmo.ru)

области здравоохранения) и технологических ограничений (ограничения существующих алгоритмов машинного обучения и других используемых методов, ограничения медицинской системы в контексте сценариев взаимодействия с пациентами, ограничения со стороны медицинских информационных систем) [5].

Основной целью развития концепции eHealth на данный момент является разработка инструментов и технологий телемедицины следующего поколения [6]. Это подразумевает преодоление географических и временных барьеров между службами здравоохранения и их пользователями, а также усовершенствование работы алгоритмов машинного обучения, которые направлены на улучшение взаимодействия между специалистами и пациентами (что подразумевает переход от экстренного оказания медицинской помощи к профилактическим мерам, а это возможно только на основе управления взаимодействием на данных).

Для актуализации сферы применения технологий ИИ достаточно упомянуть исследования, которые свидетельствуют о том, что на врачей с каждым годом возрастает офисная нагрузка (связанная с формальными протоколами оказания медицинской помощи), что влечет за собой профессиональное выгорание. В исследованиях отмечается [7, 8], что каждый пятый врач и две из пяти медсестер говорят, что из-за этого оставят свою профессию в течение двух лет. Использование чат-ботов и технологий, основанных на данных, может значительно снизить нагрузку на врачей, медсестер и других медицинских работников при сборе симптомов и прочих данных.

Стоит отметить, что на данный момент многие больницы и системы здравоохранения уже используют автоматизацию, чтобы облегчить проблему нехватки персонала и эффективно управлять и сортировать пациентов в больших масштабах. Одним из видов автоматизации являются чат-боты и связанные с ними технологии, которые могут автоматизировать процесс приема симптомов.

Хорошим примером интеграции чат-ботов в медицинскую систему является Northwell Health, система здравоохранения в Нью-Йорке, которая использует автоматические чаты, чтобы помочь своим лечащим бригадам оставаться на связи с пациентами во время их реабилитации вне больницы.

Пациенты взаимодействуют с цифровым помощником, который выходит на связь с пациентами от имени клиники. Он задает вопросы об их состоянии здоровья и лечении, а также может считывать биометрические данные с устройств. Чат-бот анализирует ответы пациентов в режиме реального времени и рекомендует пациенту, что делать дальше (корректировка лечения). Если бот определяет, что пациент нуждается в дополнительной помощи, он направляет его в клинику, переводя пациента на телемедицинский визит или на колл-центр записи на личный прием.

Несмотря на перспективность этих направлений разработки, существует ряд проблем при внедрении человеко-машинного диалога в повседневные практики взаимодействия клиник и пациентов. Одна из этих проблем – отсутствие желания со стороны пользователей взаимодействовать со сценариями, основанными на правилах (так как это означает не эмпатичный диалог, а также ощущение отсутствия возможности отклонения от сценария, заложенного создателями интерфейса). Решение этой проблемы заключается в реализации эмпатичного диалога. Это значит, что диалоговый агент должен эмулировать человеческие черты. Таким образом, первичная проблема кроется в способности поддерживать диалог на естественном языке (а не посредством систематического опроса обратившегося пациента по полному списку протокольных вопросов).

II. ПРОБЛЕМА

Первый приём пациента в медицинских организациях начинается с того, что собирается анамнез пациента, после чего на фоне собранных данных выносится предварительный диагноз, далее пациенту рекомендуется сдать необходимые анализы, чтобы подтвердить или опровергнуть предварительный диагноз. Процедура формирования анамнеза и процесс вынесения первичного диагноза требуют много временных и человеческих ресурсов со стороны квалифицированного персонала. Следовательно, если медицинское учреждение встречается с большим потоком пациентов (что зависит, например, от количества клиник в районе, и, соответственно, распределения жителей по ним), то рутинная нагрузка на врачей при первичном приеме пациентов снижает фокус внимания и, как следствие, качество лечения. При этом сбор анамнеза и постановка предварительного диагноза могут быть автоматизированы с помощью некоторой модели на базе машинного обучения, за счет делегирования ей части рутинных и аналитических процессов. Таким образом, врачи, получая данные от модели, могли бы иметь время и эмоциональный ресурс на большее внимание к пациенту, начиная прием не с шаблонных действий, а с фокусировки на проблеме.

На данный момент существует несколько готовых проприетарных решений автоматизации. Первое возможное решение в пользу экономии времени врача – экспертно-справочные медицинские системы. В данном случае речь идет о справочных системах для поиска или уточнения диагноза по заданным симптомам. Примерами таких систем являются разработка компании Mail.ru «Карта симптомов», симптомчекер от компании Webiomed, система «Консилиум», система Medai (Медаи), проект MeDiCase. Есть системы, доступ к информации в которых выдается только по логинам и паролям со стороны медицинских учреждений, и в то же время существуют открытые проекты, которые доступны обычному интернет-пользователю. В целом, существенный недостаток такого рода снижения нагрузки на врача заключается в том, что речь идет о предоставлении цифрового справочника симптомов с удобным интерфейсом, но не о уменьшении бюрократических протоколов.

Есть другой подход к попытке оптимизации первого этапа взаимодействия врача и пациента – использование для взаимодействия «клиника-пациент» примитивных диалоговых агентов, основанных на правилах и ветвлениях. Примером этого подхода является реализованный в сервисе «СберЗдоровье» чат-бот. Проведенный нами опрос (декабрь 2021 года, жители

города Санкт-Петербург, 95 респондентов) показал, что индивиды, предпочитающие обращаться в медицинские учреждения посредством дистанционного контакта на начале взаимодействия, раздражаются негибкости подобных ботов: приходя в интерфейс диалогового агента с определенной проблемой, которую необходимо решить (чаще всего – запись к врачу), пользователи оказываются закованными в рамки длинного опросного листа, прерывание взаимодействия с которым означает недостижение изначально поставленной цели.

Обнаруживается и общий недостаток этих направлений работы над оптимизацией первого контакта пациента с врачом – данные решения не поддерживают работу с естественным языком. Они представляют пользователю некий интерфейс, что, как предполагается, должно облегчить получение необходимой информации. Но вместе с тем ни один интерфейс не поддерживает обработку естественного языка, на котором привыкли разговаривать пациенты, описывая свои симптомы и болезни. Для направления, связанного с диалоговыми агентами, которые могут и должны быть интегрированы в контур информационных систем медицинских учреждений, это представляется наиболее критичной проблемой для дальнейшего развития логики дистанционных взаимодействий: если пациент сможет описать свое состояние на естественном языке, то есть не следуя по длинному прямолинейному пути опросника, а модули обработки текстовых данных, подключенные к боту, смогут верно выделить необходимые сегменты текста и проанализировать поступившие сведения, то такой контакт будет удобен и клинике, и пользователю.

Таким образом, мы поставили перед собой цель разработать языковой модуль, который бы позволял из строки текста (сообщения пользователя) собирать анамнез, а далее проанализировать полученные данные и предоставить предварительный диагноз (идея мультиклассовой классификации по симптомам, основанная на работе специфической языковой модели, фокусирующейся на симптомах). Тогда схема взаимодействия с пациентом может быть представлена следующим образом (рис. 1).

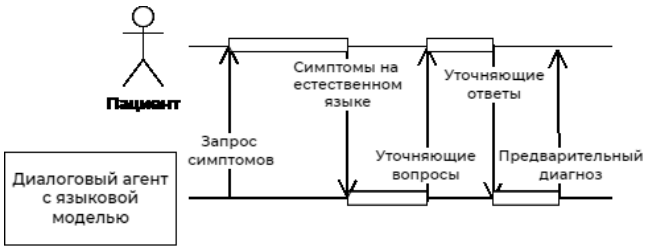

Рис. 1. Взаимодействие «пациент-клиника» при подключении языкового модуля к диалоговому агенту

Иными словами, предполагается, что разработанный модуль позволит двигаться по следующей траектории: диалоговый агент запрашивает симптомы, пациент пишет свое состояние (ощущения), описывая его на естественном языке, бот обрабатывает поступившие данные с использованием модуля, при необходимости (если модель классификации не достигает адекватного вероятностного распределения) запускается ветка алгоритма, отвечающая за уточняющие вопросы, пациент отвечает на них также на естественном языке. В итоге модель классификации предоставляет предварительный диагноз с объяснением, почему именно он наиболее вероятен.

Таким образом, назначение модуля видится в возможности на его базисе реализовать две на данный момент недоступные для специализированных диалоговых агентов функции: 1) выделение специфических именованных сущностей (симптомы); 2) классификация заболевания на основе составленного из них анамнеза.

III. ДАННЫЕ

Итоговый датасет выложен в открытом доступе на платформе Kaggle[1] и насчитывает более 5 тыс. записей.

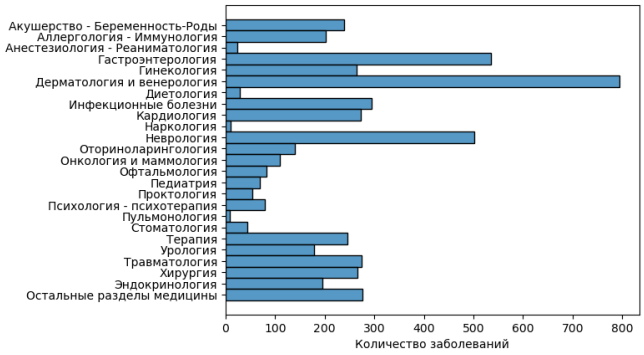

Рис. 2. Распределение категорий болезней

Изначально в качестве данных мы использовали априорные знания о болезнях и их симптомах, которые собрали с использованием веб-ресурса meduniver.com. Они были подвергнуты ручной очистке и приведены в вид «болезнь – симптомы». Всего было обработано 106 заболеваний. Слабоструктурированные данные были очищены и приведены в вид: область медицины – подтверждение (1) или отрицание (-1) симптома. На рис.2 представлено имеющееся распределение болезней по категориям. Всего было выделено 480 симптомов. Слабоструктурированные данные были очищены и приведены в вид: область медицины – подтверждение (1) или отрицание (-1) симптома. Далее были проведены первые эксперименты по мультиклассовой классификации текстов на естественном языке, в которых описано самочувствие и ощущения пациентов. Метрики качества показывали очень низкие значения качества моделей (<0.5). Это показало, что для обучения моделей необходимо собрать текстовые данные другого рода, а именно описания заболеваний со стороны пациентов в формате естественного языка. Такие данные нам удалось найти на форумах и в других открытых источниках, где цитируется прямая речь заболевших и есть маркер заболевания со стороны медицинского сообщества.

Стоит отметить, что в открытом доступе не оказалось достаточно данных с описаниями анамнезов пациентов (с использованием их прямой речи), так как правовое поле ограничивает возможности для обнародования подобной информации со стороны врачей. Поэтому было решено сгенерировать искусственные анамнезы

[1] https://www.kaggle.com/datasets/egorovm/patient-disease

пациентов с помощью инструмента GPT-3.5. На вход ChatGPT получал указание, какие симптомы ему нужно использовать в своем «рассказе о самочувствии», а какие нет. На выходе ChatGPT выдавал анамнез. Всего было сгенерировано 5563 различных анамнезов пациентов для 10 видов заболеваний: «инсульт», «ангина», «корь», «ветрянка», «цистит», «тахикардия», «отит», «анемия», «гастрит». Проверка адекватности сгенерированных описаний самочувствия проводилась вручную, с привлечением респондентов (5 медиков, случайные выборки из данных по 500 строк).

IV. ОБЩАЯ ИДЕЯ МОДУЛЯ

На рис. 3 представлена общая структура создаваемого модуля.

Рис. 3. Последовательность процессов взаимодействия с текстом на естественном языке внутри модуля

Компонента предварительной обработки текста отвечает за обработку текста сообщения пользователя. Логика ее работы практически не отличается от стандартного набора процедур (удаление спецсимволов, знаков пунктуации, далее – лемматизация слов), однако мы решили не удалять ряд стоп-слов (таких как «не», «нет» и т.п.), так как их было важно сохранить для учета отрицающихся симптомов, например, оборота речи «температуры нет».

Следующая компонента (далее в тексте - SymptomExtractor) отвечает за извлечение симптомов из сообщений пациента и, как следствие, преобразование текста в векторное представление. Для выделения симптомов была использована библиотека Spacy, а также заранее сгенерированные правила для выделения симптомов с помощью нее. Мы сгенерировали правила для чуть больше 5000 симптомов. Отрицания симптомов обрабатывались с помощью библиотеки negspacy. Таким образом, получается векторное представление симптомов, которое можно использовать для обучения модели и предсказании болезни. Таким образом, этот подмодуль содержит набор правил, по которым выделяются симптомы из сообщений пациента с присвоением статуса (да, нет, нет информации, непонятно). Компонента классификации заболевания отвечает за предсказания предварительного диагноза по векторному представлению, полученному на предыдущем шаге.

Последнее, что делает модуль – предоставляет объяснение (рис. 4), почему модель классификации выдала именно такое предсказание, и какие симптомы на это больше всего повлияли. Полученная информация может передаваться в информационную систему для принятия дальнейших решений (срочность приема, «второе мнение» лечащему врачу).

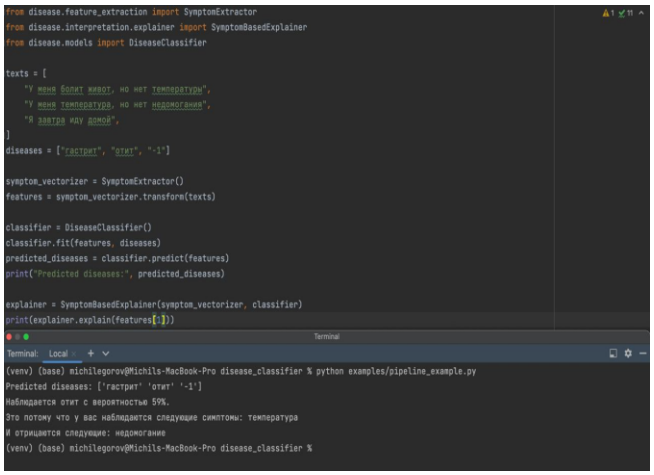

Рис. 4. Интерпретируемое суждение со стороны модуля относительно возможного диагноза пациента

Как видно из скриншота модель выводит вероятность предполагаемого диагноза, а также объяснение по каким симптомам сделан вывод: детектированные симптомы и отрицающиеся симптомы выводятся сепарированными строками, для большей наглядности.

V. МЕТОД КЛАССИФИКАЦИИ ЗАБОЛЕВАНИЙ

Подмодуль для определения болезней был разработан как ключевой компонент модуля, его задачей было использовать выделенные симптомы для классификации и определения вероятной болезни. Основной идеей была обработка векторов симптомов и преобразование их в соответствующие диагнозы. Необходимо было учесть, что в сфере здравоохранения важным признаком моделей является интерпретируемость результатов. Исходя из этого, было решено обучить и провалидировать два подхода: классический подход с использованием логистической регрессии и AutoML фреймворк FEDOT (который подходил за счет его объясняющей механики).

FEDOT был выбран в качестве автоматизированного решения для подбора оптимальной модели. Он позволяет оптимизировать как структуру модели, так и параметры используемых алгоритмов. Это позволяет автоматически выбирать наиболее подходящую модель для каждого конкретного набора данных и задачи. Однако, несмотря на эффективность AutoML, эти модели могут быть сложными и непрозрачными для интерпретации.

Логистическая регрессия является более простой моделью, которую можно легко интерпретировать. Коэффициенты модели логистической регрессии могут быть напрямую связаны с вероятностями классов и использованы для определения важности каждого симптома при определении болезни. Это делает логистическую регрессию привлекательной для использования в клинической практике, где важно не только точно классифицировать состояние пациента, но и объяснить причины такого предсказания.

Эти варианты были протестированы с использованием сгенерированных случаев пациентов.

Сначала была проведена предварительная подготовка текста: очистка от знаков препинания, приведение к нижнему регистру, нормализация слов и удаление стоп-слов. Важно отметить, что из процесса удаления стоп-слов было исключено удаление слова «нет» и похожие на него конструкции, чтобы не потерять факт отрицания симптома. Далее для каждого случая извлекались симптомы пациентов.

Ниже приведены результаты тестирования моделей.

Таблица. Сравнение результатов тестирования моделей

Модель	Векторизатор	Recall	Precision	F1-score
Logistic Regression	Tf-Ifd Vectorizer	0.19	0.18	0.18
Logistic Regression	SymptomExtractor	0.15	0.23	0.16
AutoML FEDOT	SymptomExtractor	0.18	0.14	0.15

Из таблицы видно, что все модели достаточно много ошибаются, однако если сравнить SymptomExtractor с базовым Tf-Idf, который переобучился на n-gram-ах, то становится очевидным, что подход с извлечением симптомов себя оправдывает, увеличение точности в этом случае возможно посредством расширения обучающих данных (рис. 5).

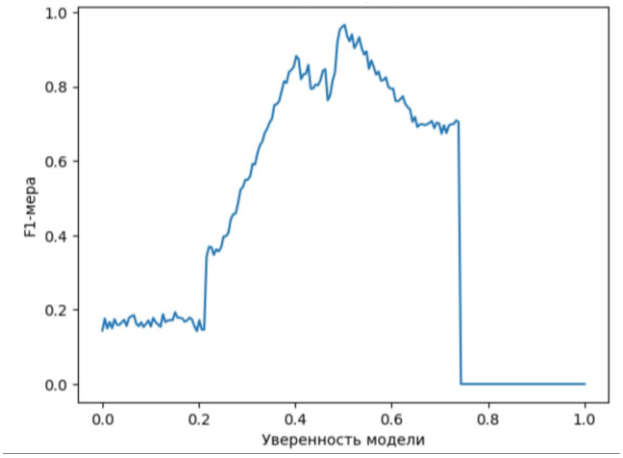

Рис. 5. Зависимость F1-меры от уверенности модели

Несмотря на то, что F1-мера ниже, подход с использованием SymptomExtractor показал наивысшую Precision, что важно для работы с диагнозами.

Подход с AutoML FEDOT не показал хороших результатов из-за многообразия входных признаков и малого количества случаев, но дальнейшее сжатие признаков и увеличение обучающего датасета позволит использовать подходы основанные на AutoML.

Еще одним преимуществом использования SymptomExtractor является то, что в весах LogisticRegression не случайные n-gramm-ы, а настоящие симптомы. Веса, присвоенные каждому симптому, отражают важность этого симптома для определения заболевания. Таким образом, врачи и другие медицинские специалисты могут анализировать веса и понимать, какие симптомы являются наиболее информативными для определения конкретного заболевания. Более того, логистическая регрессия позволяет учитывать наличие, отсутствие и отрицание симптомов, что делает результаты еще более точными и надежными. Далее были рассмотрены веса каждого из симптомов для определения болезни (рис.6).

В целом, полученные результаты вполне соответствуют тому, что можно было бы ожидать, определяя заболевание, например, с использованием симптомочекеров, упомянутых выше. Однако можно заметить несколько интересных контринтуитивных наблюдений, выявленных моделью:

1) Ангина. Самым важным симптомом, указывающим на наличие ангины, с точки зрения статистической модели является анемия. Это может показаться нелогичным, так как анемия обычно не ассоциируется с ангиной. Однако этот результат отражает то, что анемия может ухудшать симптомы ангины или увеличивать вероятность ее развития.

Рис. 6. Интерпретируемое суждение со стороны модуля относительно возможного диагноза пациента

2) Гастрит. Одним из наиболее важных симптомов, указывающих на наличие гастрита, явилось онемение. Здесь также важно отметить, что на этот симптом вряд ли можно обратить внимание при «стандартном» фиксировании симптомов (в частности потому, что вопрос об онемении не входит в перечень протокола первичного опроса пациента).
3) Тахикардия. Боль в мышцах является одним из наиболее важных симптомов, указывающих на отсутствие тахикардии, это может означать, что наличие этого симптома уменьшает вероятность наличия у пациента тахикардии (либо, что этот симптом слишком распыляет гипотезы о присутствующем заболевании).

В приведенных диаграммах можно увидеть и другие закономерности. Таким образом, эти результаты демонстрируют важность использования статистических моделей для анализа данных о здоровье, так как они могут обнаруживать неожиданные связи между симптомами и заболеваниями, что может помочь врачам лучше понимать, как диагностировать и лечить своих пациентов.

При малой уверенности модели в своем решении, необходимо спросить у пациента дополнительную информацию или перенаправить в медицинское учреждение. При достаточно уверенных прогнозах (>0.6) модель показывает F1-меру = 0.89.

VI. ЗАКЛЮЧЕНИЕ

Предпосылкой для проведения данного исследования стал опыт создания мультимодального интеллектуального помощника для автоматизации процесса приема пациентов и оказания первичной медицинской помощи. В ходе экспериментов было выяснено, что готовый интерфейс чат-бота, который предполагает наличие сценариев взаимодействия с пациентом, которые заложены разработчиками, не очень применим к реальной ситуации, когда медицинское учреждение запускает на своем сайте или внутри мобильного приложения бота. Это происходит в связи с тем, что протоколы разных клиник при взаимодействии с пациентом могут отличаться, как и техническое задание со стороны администрации. Все зависит от основной целевой функции диалогового агента: разгружает ли он регистратуру, может ли он быть бесшовно подключен к информационной системе, могут ли передаваться данные из него на компьютер принимающего врача и т.п. Таким образом, мы пришли к выводу, что разработчики чат-ботов на стороне клиники должны иметь языковой модуль, функционал которого может быть использован частично или полностью, но, главное, результаты работы каждого подмодуля интегрируются в продуманный на стороне медицинского учреждения сценарий работы диалогового агента. Таким образом, наша цель была создать многокомпонентный модуль, который мог бы выделять симптомы и их отрицания, классифицировать на этом фоне заболевания, а затем выносить интерпретируемое суждение. В настоящее время мы работаем над улучшением значений метрик качества и планируем в итоге измерять качество модуля за счет привлечения медицинских экспертов для тестирования. Нам видится, что наборы данных должны быть дополнены новыми случаями для увеличения точности моделей. Однако мы считаем, что модуль показывает применимость на практике. Наборы данных и библиотека находятся в свободном доступе на GitHub[2].

Благодарности

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».

[2] https://github.com/NIRMA-PATIENT-INTAKE/disease

Библиография

[1] Cardoso A. R. C., Bento B. A. C. Evolution, applicability, new challenges and opportunities in Telemedicine // 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2016. P. 1-6.

[2] Niazkhani Z., Pirnejad H., Khazaee P. R. The impact of health information technology on organ transplant care: a systematic review // International Journal of Medical Informatics. 2017. Vol. 100. P. 95-107.

[3] Peters T. E. Transformational impact of health information technology on the clinical practice of child and adolescent psychiatry // Child and Adolescent Psychiatric Clinics. 2017. Vol. 26. № 1. P. 55-66.

[4] Sanderson J. A Human Rights Framework for Intellectual Property // Innovation and Access to Medicine. 2016. Vol. 11. № 2. P. 149-150.

[5] Isetta V. et al. A Bayesian cost-effectiveness analysis of a telemedicine-based strategy for the management of sleep apnoea: a multicentre randomised controlled trial // Thorax. 2015. Vol. 70. № 11. P. 1054-1061.

[6] Vimarlund V., Le Rouge C. Barriers and opportunities to the widespread adoption of telemedicine: a bi-country evaluation // Studies in health technology and informatics. 2013. Vol. 192. P. 933-933.

[7] Sinsky C. A. et al. 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.

[8] Sinsky C. et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties // Annals of internal medicine. 2016. Vol. 165. № 11. P. 753-760.

Чижик Анна Владимировна, к. культ., старший научный сотрудник Центра технологий электронного правительства Института дизайна и урбанистики, Университет ИТМО (http://egov.itmo.ru/), Санкт-Петербург, email: afrancuzova@mail.ru, elibrary.ru: authorid=708001, scopus.com: authorId=57222136821, ORCID: orcidID=0000-0002-4523-5167

Егоров Мичил Прокопьевич, старший лаборант Центра технологий электронного правительства Института дизайна и урбанистики Университета ИТМО, Санкт-Петербург (https://egov.itmo.ru/), email: egorovm@niuitmo.ru, scopus.com: authorId=57226072648, ORCID: orcidID= 0000-0002-0125-7540

Creating a Language Module for Predicting a Patient's Preliminary Diagnosis for Interaction through a Dialog Agent

Anna V. Chizhik, Michil P. Egorov

Abstract— **In recent years, conversational agents based on artificial intelligence are considered as a promising method of primary interaction with a patient when he visits a clinic. Indeed, this helps to relieve the registry, as well as optimize the flow of patients within the medical institution. It is worth noting that the customer experience that individuals have gained through the active use of the online environment to solve everyday tasks motivates them to implement dialog agents in such a way that the interaction is quick and convenient. It is due to this that motivation is created for the further use of this communication channel. However, chatbots in Russian currently do a poor job of interpreting text data received from the user. Thus, it is a difficult task to isolate symptoms from a user-entered string, and in connection with this there is a further problem of text classification.. The paper is devoted to the description of the process of designing and developing a language model for predicting a patient's preliminary diagnosis when interacting using a dialog agent. A priori knowledge about diseases and their symptoms, semi-structured data from disease general discussion forums, and a generated set of texts using ChatGPT were selected as data for model training and algorithm validation. The article describes the general idea of the created library, reveals the topic of disease classification, and analyzes the quality metrics of the developed models.**

Keywords— **text classification, logistic regression, language model, conversational agents, eHealth.**

REFERENCES

[1] Cardoso A. R. C., Bento B. A. C. Evolution, applicability, new challenges and opportunities in Telemedicine // 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2016. P. 1-6.

[2] Niazkhani Z., Pirnejad H., Khazaee P. R. The impact of health information technology on organ transplant care: a systematic review // International Journal of Medical Informatics. 2017. Vol. 100. P. 95-107.

[3] Peters T. E. Transformational impact of health information technology on the clinical practice of child and adolescent psychiatry // Child and Adolescent Psychiatric Clinics. 2017. Vol. 26. № 1. P. 55-66.

[4] Sanderson J. A Human Rights Framework for Intellectual Property // Innovation and Access to Medicine. 2016. Vol. 11. № 2. P. 149-150.

[5] Isetta V. et al. A Bayesian cost-effectiveness analysis of a telemedicine-based strategy for the management of sleep apnoea: a multicentre randomised controlled trial // Thorax. 2015. Vol. 70. № 11. P. 1054-1061.

[6] Vimarlund V., Le Rouge C. Barriers and opportunities to the widespread adoption of telemedicine: a bi-country evaluation // Studies in health technology and informatics. 2013. Vol. 192. P. 933-933.

[7] Sinsky C. A. et al. 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.

[8] Sinsky C. et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties // Annals of internal medicine. 2016. Vol. 165. № 11. P. 753-760.

Anna V. Chizhik, Ph.D in Cultural Studies, Senior Researcher of E-Governance Center, Institute of Design and Urban Studies, ITMO University (http://egov.itmo.ru/), Saint-Petersburg, email: afrancuzova@mail.ru, elibrary.ru: authorid=708001, scopus.com: authorId=57222136821, ORCID: orcidID=0000-0002-4523-5167

Michil P. Egorov, Senior Assistant of E-Governance Center, Institute of Design and Urban Studies, ITMO University (http://egov.itmo.ru/), Saint-Petersburg, email: egorovm@niuitmo.ru, scopus.com: authorId=57226072648, ORCID: orcidID= 0000-0002-0125-7540