

УДК 004.89:61

10.25209/2079-3316-2024-15-4-153-181



Извлечение симптомов и автоматическое предсказание диагноза по медицинским клиническим записям

Юрий Петрович **Сердюк**¹, Наталья Александровна **Власова**²,
Седа Рубеновна **Момот**³

¹⁻³Институт программных систем им. А. К. Айламазяна РАН, Веськово, Россия

Аннотация. В статье представлена система, которая извлекает симптомы заболеваний из медицинских клинических записей (текстов на естественном русском языке) и автоматически предсказывает по ним диагноз в виде наименования заболевания и его кода в соответствии со справочником МКБ-10. Система ограничена предметной областью из 6 пульмонологических заболеваний (хроническая обструктивная болезнь легких, пневмония, бронхиальная астма и др.) и COVID-19.

Извлечение симптомов реализовано с помощью нескольких нейронных сетей, выделяющих отдельные медицинские сущности и связи между ними. Предсказание диагноза также реализовано в виде классификатора на основе нейронной сети. Для обучения извлечению симптомов создан аннотированный корпус предложений. Описаны принципы и правила разметки симптомов. Представлен корпус текстов для обучения классификатора предсказанию диагнозов.

Приведены оценки точности при тестировании обеих подсистем. Точность предсказания диагноза на данной предметной области составила 88,5%. Даны сравнения с аналогичными работами по извлечению симптомов из текстов на различных языках, а также по автоматическому предсказанию диагнозов, включая системы типа ChatGPT.

Ключевые слова и фразы: системы поддержки принятия врачебных решений, извлечение симптомов, автоматическое предсказание диагнозов, BERT-модели, системы на основе ChatGPT.

Для цитирования: Сердюк Ю. П., Власова Н. А., Момот С. Р. *Извлечение симптомов и автоматическое предсказание диагноза по медицинским клиническим записям* // Программные системы: теория и приложения. 2024. Т. 15. № 4(63). С. 153–181. https://psta.psir.ru/read/psta2024_4_153-181.pdf

Введение

Одним из главных трендов развития современных систем здравоохранения является применение высокотехнологичных решений с использованием искусственного интеллекта (ИИ). В первую очередь методы ИИ находят свое применение в так называемых системах поддержки принятия клинических (врачебных) решений (СППКР/СППВР) [1]. Эти системы прошли долгий путь развития начиная с 1990-х годов, когда они строились на основе различного рода экспертных систем, базировавшихся, в свою очередь, на знаниях в виде правил или семантических сетей (фреймов), логических языках программирования типа PROLOG и других средствах [2].

На настоящий момент подавляющее число систем, работающих в области здравоохранения и использующих элементы ИИ, ориентировано на обработку медицинских изображений. Однако в последние годы существенный прогресс в области ИИ связан с применением нейронных сетей для обработки текстовой медицинской информации, включая «большие языковые модели» (Large Language Models — LLM), а также системы типа ChatGPT, которые позволяют намного более эффективно, чем ранее, решать задачи обработки (медицинских) текстов на естественном языке (ЕЯ) [3–5].

Существует несколько способов использования нейронных сетей и, в частности, LLM для обработки медицинских текстов. Часто нейронные сети обучают для извлечения значимой информации из текстов — в первую очередь симптомов и всего, что с ними связано [6, 7]. С другой стороны, системы на базе LLM применяют для прямой постановки диагноза по записям в (электронной) медицинской карте пациента [4, 8]. Также нейронные модели используются, например, для извлечения отзывов пациентов на применение тех или иных лекарств [9].

Упомянутые выше системы обладают рядом недостатков (подробный сравнительный обзор этих систем дан в разделе 3 «Обзор смежных работ»). В частности, в работе [7] симптомы извлекаются не в структурированном виде (см. раздел 1), а в форме «упоминаний симптомов», и дальнейшая постановка диагноза осуществляется на основе множества решающих правил. Семейства решающих правил обладают теми известными недостатками, что они трудоемки в разработке, у них обычно отсутствует полнота охвата возможных симптомов и их комбинаций, а при увеличении количества таких правил становится трудно следить за их непротиворечивостью.

С другой стороны, доказано, что нейронные сети способны смоделировать (вычислить) любую (предсказательную) функцию [21, Гл. 4].

В работе [4] предлагается подход к обучению нейронной сети простой классификации текстов (электронных записей о пациенте) на 265 классов, соответствующих заболеваниям из Международной классификации болезней версии 10 (МКБ-10). В данной системе отсутствует извлечение симптомов как таковое, что, в частности, делает невозможным какое-либо объяснение поставленного диагноза.

Системы типа ChatGPT вообще оказались практически непригодными в задачах автоматической постановки диагноза — диагностическая система на базе GPT-3.5 допустила ошибки в 83% педиатрических случаев, а система на базе GPT-4 правильно диагностировала только 39% сложных медицинских случаев у взрослых и детей [10].

В данной статье описывается система, в которой предпринята попытка устранения недостатков, присутствующих в упомянутых подходах. Постановка диагноза врачом является сложным многоэтапным процессом, в результате которого формируется ответ обычно состоящий из нескольких структурных частей:

- (1) основного диагноза,
- (2) осложнений основного заболевания,
- (3) сопутствующих заболеваний.

В настоящей работе моделируются два этапа процесса постановки диагноза:

- (1) извлечения симптомов из (электронной) медицинской карты пациента,
- (2) формулировки гипотезы о заболевании в виде его наименования и кода МКБ-10.

Хотя в современных медицинских информационных системах (МИС) информация о пациенте вводится через заранее подготовленные шаблоны, что делает ее частично структурированной, некоторая ее часть (иногда значительная и включающая в себя важные диагностические признаки) остается в свободном неструктурированном виде. Представляемая система способна извлекать симптомы из любого текста (структурированного и неструктурированного), представляя их в едином формате.

В первую очередь извлекаются те симптомы заболевания, которые наиболее важны и специфичны для заданной группы заболеваний. Кроме того, развитые подсистемы извлечения симптомов из неструктурированного текста позволят уменьшить трудоемкость разработки и сопровождения МИС, в частности, программирования и репрограммирования входных шаблонов, оставив только стандартные шаблоны для всех или определенных групп заболеваний.

Шаг извлечения симптомов из имеющихся медицинских документов, относящихся к данному пациенту, и предъявление их врачу можно рассматривать как частичное объяснение выставленного впоследствии диагноза. Для получения максимально значимых практических результатов представляемая система ограничена предметной областью из шести пульмонологических заболеваний и COVID-19.

Результатом работы системы являются

- (1) список симптомов в структурированном виде — симптом и его значение,
- (2) гипотеза о диагнозе в виде названия заболевания, его кода по классификации МКБ-10, процента вероятности данного диагноза.

В случае, когда система не может с достаточной степенью уверенности предсказать заболевание, об этом выдается специальное сообщение.

Таким образом, предлагаемую систему можно использовать в качестве помощника врача, систематизирующего и представляющего в едином формате информацию, собранную о пациенте, и предлагающего гипотезу о вероятном заболевании.

Статья построена следующим образом:

В разделе 1 «Извлечение диагностических признаков» описан этап извлечения симптомов, которые мы расширяем до более общего понятия «диагностические признаки». Представлен аннотированный корпус предложений для обучения нейронной сети извлечению диагностических признаков, описаны принципы и правила их разметки.

В разделе 2 «Постановка диагноза» описан классификатор на основе нейронной сети, который осуществляет постановку диагноза по извлеченным диагностическим признакам. Представлен корпус из 140 документов, использованных для обучения классификатора.

В разделе 3 даны сравнения представляемой работы с аналогичными работами по извлечению симптомов, как с отечественными, так и с зарубежными, включая системы по извлечению симптомов COVID-19. Аналогично, приведены сравнительные сведения о системах автоматической постановки диагнозов, включая системы, разработанные в СберМедИИ, и системы на основе ChatGPT.

В заключительном разделе 4 сделаны краткие выводы о проделанной работе и обозначены пути дальнейшего совершенствования представленной системы.

1. Извлечение диагностических признаков

В рамках описываемого в данной статье подхода одним из этапов процесса диагностики является извлечение из (электронной) медицинской карты пациента, включающей обычно неструктурированную часть, всей необходимой для этого информации. Соответственно, ниже мы применяем термин «диагностический признак», понимая под ним не только отклонения и нарушения работы организма пациента (симптомы заболевания), но и любую другую информацию, которая может быть использована врачом (а также автоматической системой) для постановки диагноза. Иногда для краткости изложения мы будем использовать термин «симптом», понимая под ним диагностический признак.

В качестве диагностических признаков обычно рассматриваются

- (1) симптомы определенных заболеваний (малопродуктивный кашель, дизурия, менингеальные знаки) — такие симптомы мы называем симптомами-отклонениями,
- (2) описание состояния и характеристики органов или частей тела (кожные покровы бледно-розовые, сердечные тоны ясные, зев спокоен) — такие описания мы называем симптомами-характеристиками,
- (3) выражения с числовыми данными (температура $38,6^{\circ}\text{C}$, частота дыхательных движений 21–22 в минуту, сатурация 98%) — такие выражения называются показателями,
- (4) сведения об анамнезах пациента и его общем состоянии.

К результату работы подсистемы извлечения диагностических признаков предъявляются два основных требования:

- (1) понятность, наглядность и естественность представления выделенных из документа признаков (возможно, с ориентацией на определенную группу заболеваний),
- (2) возможность их эффективного использования при постановке диагноза автоматической системой.

В частности, первое требование позволяет использовать выделенные признаки как часть объяснения поставленного системой диагноза.

Возможны несколько подходов к аннотированию симптомов/диагностических признаков со структурной точки зрения:

- (1) неструктурное аннотирование, когда выделяются выражения, представляющие собой «упоминания симптомов»:
 - «малопродуктивный кашель»,
 - «миндалины гиперемированы, гипертрофированы»,

«грудная клетка правильной формы»,
«одышка при повышении физической нагрузки»;

(2) разбивка диагностического признака на две части — непосредственно на признак и на его значение:

[кашель; *малопродуктивный*]
[миндалины; *гиперемированы, гипертрофированы*]
[грудная клетка; *правильной формы*]
[одышка; *при повышении физической нагрузки*]

(3) выделение семантических медицинских сущностей с последующим конструированием из них диагностических признаков.

Первый подход [11] обладает тем недостатком, что различных упоминаний симптомов имеется очень большое количество, и потому для них трудно построить систему, прогнозирующую диагноз. В этом случае иногда прибегают к упрощению понятия «симптом/диагностический признак», выбрасывая из упоминания симптома некоторые его части, например, в выражении «одышка при повышении физической нагрузки» оставляют только слово «одышка» [7], что, естественно, намного снижает возможности системы по постановке правильного диагноза.

Третий подход состоит в том, что первоначально в тексте выделяются такие сущности как «заболевание (нарушение)», «симптом», «часть тела», «физиологический процесс», «степень тяжести (интенсивность)», «характеристика» и некоторые другие ¹, а далее из них собираются более крупные структурные единицы — диагностические признаки, на основе которых уже ставится диагноз. Недостаток этого подхода состоит в том, что очень трудно сформулировать достаточно точные правила относящие слова и словосочетания к той или иной семантической категории — примером является выражение «затруднение носового дыхания», что делает процесс аннотирования неоднозначным и очень сложным.

В нашей работе мы применяем второй подход, в котором диагностический признак считается состоящим из двух частей — непосредственно признака и его значения. Этот подход моделирует медицинскую практику врачей, когда в качестве результата обследования пациента перед этапом постановки диагноза формируется карта извлеченных симптомов (признаков). На рисунке 1 показан пример фрагмента карты обследования пациента, формируемой в одной из МИС с использованием шаблонов.

¹см., например, описание схем разметки на странице *аннотированного корпуса клинических текстов, созданного совместно ИСА РАН и ИЦЗД.*⁽⁶⁾

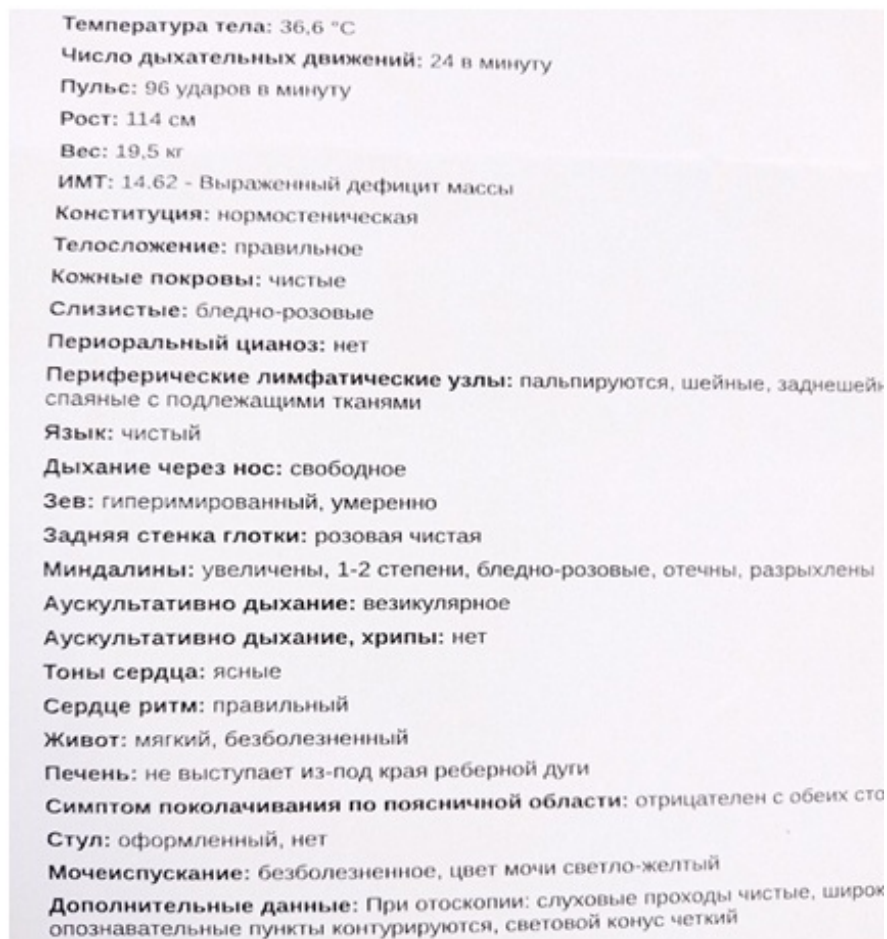


Рисунок 1. Пример результирующей карты обследования пациента.

Представляемая система позволяет строить аналогичную карту из любого свободного, неструктурированного текста. Кроме того, сама разметка признаков и их значений более проста по сравнению с другими подходами к аннотированию медицинских сущностей (см. Принцип I ниже).

Далее выделенные признаки оформляются в виде входов классификатора, на которые подаются соответствующие им значения. Общая схема разметки, выработанная еще в работе [11], состоит из

- (1) **принципов разметки**, относящихся ко всем аннотируемым выражениям,
- (2) **правил разметки**, относящихся к большим группам выражений, объединенных некоторым общим свойством и
- (3) **специальных случаев** — правил разметки, относящихся обычно к небольшим по размеру, специальным группам выражений.

Ниже дается сокращенное описание схемы разметки, включающее в себя принципы и основные правила разметки, а также несколько специальных случаев.

Отметим, что базовой единицей аннотированного корпуса для извлечения диагностических признаков является предложение, т.е. признаки и соответствующие им значения выделяются человеком-аннотатором в рамках одного предложения.

Поскольку признак и его значение могут находиться в разных местах предложения, то для их связывания используют специальные программные средства (визуальные средства аннотирования, в нашем случае это система BRAT, <https://brat.nlplab.org/>), которые позволяют соединять различные сущности (в нашем случае признаки и их значения) с помощью направленных стрелок, см. рисунок 2.

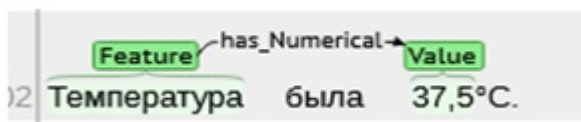


Рисунок 2. Признак со значением.

1.1. Принципы разметки

Поскольку аннотированием медицинских документов занимаются врачи, то основными критериями, предъявляемыми к принципам разметки, являются их простота, понятность, наглядность и естественность представления признаков и их значений как результата аннотирования.

Подход к выделению из выражения признака и его значения основан на **двух принципах**, представляемых ниже.

Далее, в примерах разметки полученный результат представляется в виде пары

[признак; значение]

возможно, с указанием типа связи между компонентами пары и дополнительными комментариями.

В такой паре выделенные признак и значение представляются в той лексической форме, в которой они встречаются в конкретном аннотируемом тексте.

«Перенес несколько эпизодов ОРИ с малопродуктивным кашлем.» — [кашлем; малопродуктивным].

Принцип I (базовый). В качестве признака (первого члена пары) выделяется структурно главная часть выражения, а в качестве значения — его зависимая (подчиненная) часть.

Обычно главной частью выражения является некоторое нарушение/отклонение или указание на часть тела, или физиологический процесс (выраженные одним или несколькими словами), а подчиненной частью — характеристики этого нарушения или части тела.

«Ярко выраженная гиперемия на щеках и шее.» — [гиперемия; ярко выраженная на щеках и шее]

«Ребенок поступил в отделение с жалобами на отеки Квинке при погрешности в диете.» — [отеки Квинке; при погрешности в диете]

«При поступлении: кожные покровы бледно-розовые, суховатые.» — [кожные покровы; бледно-розовые, суховатые]

«В лёгких дыхание везикулярное, равномерно проводится во все отделы.» — [дыхание; везикулярное, равномерно проводится во все отделы]

В некоторых случаях признаки могут не иметь характеристик. Правила разметки таких выражений представлены ниже в разделе 1.2.1.

Принцип II. В описаниях клинических осмотров и анамнезов некоторые признаки являются часто встречающимися, стандартными и играющими особую роль при постановке диагноза.

Такие признаки врачи обычно хотят видеть в каждой выдаче подсистемы извлечения диагностических признаков. Обычно состав такой группы признаков является специфичным для определенной группы заболеваний, и чаще всего в нее входят некоторые части тела и физиологические процессы. Например, состав группы выделенных признаков, связанных с аллергическими и пульмонологическими заболеваниями, может быть таким:

- (1) дыхание,
- (2) кожа (кожные покровы),
- (3) живот,
- (4) язык,
- (5) слизистые оболочки.

Принцип разметки выражений, включающих такого рода признаки состоит в том, что независимо от того, в виде какой части выражения они находятся — главной или зависимой, эти признаки всегда отмечаются в виде первого члена пары.

«С возраста 5-6 лет отмечается сухость кожи.» — [кожи; сухость с возраста 5-6 лет].

Следует отметить, что возможны случаи, когда в одно выражение могут входить несколько признаков из выделенной группы. В такой ситуации в качестве первого члена пары может быть выбран любой такой признак.

1.2. Правила разметки

Правила разметки уточняют принципы аннотирования I и II главных групп диагностических признаков:

- (1) признаков без характеристик,
- (2) признаков с характеристиками,
- (3) показателей,
- (4) анамнезов и сведений об общем состоянии пациента.

Ниже описываются основные правила разметки этих групп выражений с комментариями об их технической реализации в системе BRAT.

1.2.1. Признаки без характеристик

Одиночные, без дополнительных характеристик диагностические признаки являются простейшими признаками, встречающимися в медицинских текстах.

«У нее были типичные для коронавируса симптомы — **одышка**, **слабость**. С возраста 3-х лет у мальчика проявления полиноза: затруднение носового дыхания, симптомы конъюнктивита, **отеки Квинке**.»

Если в предложении говорится о наличии такого признака, то значением такого признака считается логическая константа **true**, если же об отсутствии — константа **false**. По этой причине такого рода признаки носят название булевских признаков. Для их представления ниже используются пары вида [признак; булевское значение].

«У нее были типичные для коронавируса симптомы — **одышка**, **слабость**.» — [одышка; true], [слабость; true]

«Аускультативно дыхание с жестким оттенком, **хрипов нет**.» — [хрипов; false]

Технически при разметке в системе BRAT булевские признаки отмечаются в виде сущностей типа Feature дополнительной установкой атрибута Negation для признаков со значением **false**.

Чаще всего такого рода признаки состоят из одного слова. Однако возможны случаи, когда они состоят из нескольких слов, обычно являющихся так называемыми «устойчивыми словосочетаниями» — часто встречающимися в медицинских текстах фразами, используемыми в качестве шаблонов. Примерами таких фраз являются «отеки Квинке», «сотрясение головного мозга», «вздутие живота» и т.п.

«*Менингеальных знаков на момент осмотра нет.*» — [менингеальных знаков; *false*]

«*Ведущий симптом в такой ситуации — диарея, затем по распространенности следуют вздутие живота, спастические боли, тошнота и рвота.*» — [вздутие живота; *true*], [спастические боли; *true*]

В последнем предложении выражение «спастические боли» потенциально могло бы быть разбито на признак и значение, но, тем не менее, оно должно быть отнесено к булевским признакам. В данном и аналогичных случаях, включая случаи для признаков с характеристиками, аннотация проводится согласно следующим правилам.

Правило 1.2.1 (а). В выражениях, в которых симптомы-отклонения выражены в абстрактной форме и требуют уточнения места их проявления — *боль, искривление, деформация, стеснение, изменение* и т.п., эти симптомы вместе с указанием на место проявления отмечаются как булевские.

«*Изменений в легких нет.*» — [изменений в легких; *false*]

Правило 1.2.1 (б). Симптомы-отклонения, однозначно связанные с определенной частью тела или органом, также аннотируются в виде булевского признака со включением части тела или органа в сам признак («*першение в горле*», «*першение по задней части глотки*»).

Специфика разметки выражений, в которых говорится об отсутствии какого-либо признака, имеющего некоторые характеристики, отражена в правилах следующего раздела.

1.2.2. Признаки с характеристиками

Признаки с характеристиками являются самой многочисленной группой размечаемых выражений. Структурно такие выражения состоят из двух частей, где первая часть представляет собой либо

- (1) нарушение/отклонение в работе организма, либо
- (2) указание на часть тела или
- (3) физиологический процесс,

а вторая часть содержит одну или несколько характеристик того, что указано в первой части.

«С этого времени у ребенка одышка появилась и на фоне физической нагрузки.» — [одышка; на фоне физической нагрузки]

«Живот мягкий, доступен глубокой пальпации.» — [живот; мягкий, доступен глубокой пальпации]

«Стул оформленный, склонен к запорам.» — [стул; оформленный, склонен к запорам]

К характеристикам относятся те части выражения, которые, по мнению аннотатора, описывают

- степень выраженности нарушения/отклонения («*сильная головная боль*»),
- время проявления («*заложенность носа в ночной период*», «*покашливание в утреннее время*»),
- условие проявления («*риноконъюнктивальный синдром при контакте с бытовыми аллергенами*»),
- состояние части тела («*кожные покровы чистые, бледные, периорбитальные тени*») и др.

Поскольку для признака может быть указано в тексте несколько характеристик, расположенных в разных частях предложения, то в системе визуальной разметки BRAT от одного признака может быть проведено несколько стрелок к этим характеристикам.

Аналогично, в выражении может быть перечислено несколько признаков, имеющих одно и то же значение. В этом случае, в системе BRAT проводятся стрелки от каждого признака к этому значению. В частности, для выражения «*стул, диурез не нарушены*» результатом разметки будут пары [стул; не нарушены], [диурез; не нарушены].

Аналогично случаю отсутствия (отрицания) признаков без характеристик, в медицинских текстах могут встречаться выражения об отсутствии признаков с характеристиками (выражения с отрицанием) — «*сухого кашля не наблюдается*». В таких ситуациях действует правило согласно которому характеристики считаются принадлежащими главной части выражения, которая становится булевским признаком со значением **false** — [сухого кашля; false].

1.2.3. Показатели

Под показателями понимаются результаты простейших инструментальных исследований, которые проводятся при первичном и последующем осмотрах пациента — температура тела, частота дыхательных движений, частота сердечных сокращений (пульс), уровень оксигенации (сатурации)

и др. Обычно, такие признаки имеют числовые значения и размечаются как признаки с характеристиками (см. рисунок 2):

[частота дыхательных движений; 21–22]

[SpO₂; 98]

[артериальное давление; 110/80]

В эту категорию попадают и другие результаты инструментальных исследований и анализов — количество эритроцитов в крови, размеры и характеристики внутренних органов, определяемые при ультразвуковом или рентгенологическом исследованиях и т.д. Единицы измерения значения того или иного показателя в саму разметку в нашем случае не включаются.

1.2.4. Анамнезы и общие сведения о пациенте

Выражения, которые описывают анамнез, эпидемиологический статус и общее состояние пациента, рассматриваются в качестве отдельных групп признаков и размечаются с помощью специальных атрибутов в системе BRAT.

Примерами таких выражений являются: «*хронические болезни отрицает*», «*вернулся из командировки в Германию 3 дня назад*», «*состояние средней тяжести*».

В текущей разметке такого рода выражения разбиты на 3 группы:

- (1) данные анамнезов жизни и болезни,
- (2) эпидемиологический статус (анамнез),
- (3) сведения об общем состоянии пациента.

Такие выражения отмечаются целиком, и им присваиваются атрибуты AofL (anamnesis of life), EpA (epidemiological anamnesis), CS (common state), соответственно, в системе BRAT.

В данные группы выражений могут входить разные сведения соответствующей направленности. В частности, в группу анамнеза могут входить сведения о наследственности пациента, хронических и перенесенных заболеваниях и др. Далее, в классификаторе болезней для такого рода выражений предусмотрены три вида входов — AofL, EpA и CS, на которые подаются извлеченные выражения соответствующих типов.

1.3. Специальные случаи

В данном разделе рассматриваются несколько специальных групп выражений и правила их разметки.

1.3.1. Заболевание как симптом

Иногда встречаются выражения, в которых упоминаются заболевания в качестве симптомов других заболеваний. Так, например, выявлено,

что вирус COVID-19 может вызывать конъюнктивит разного вида. Поэтому в соответствующих текстах он должен отмечаться в виде признака. Такого рода упоминания заболеваний аннотируются стандартным способом с разбивкой на признак и значение (если последнее присутствует в выражении). Так, отдельное слово "конъюнктивит" может отмечаться в виде, как обычно, булевского признака, а в предложении «У пациента наблюдалась выраженная диарея, плохо поддающаяся лечению» разметка соответствующего фрагмента будет иметь вид — [диарея; *выраженная*].

1.3.2. Выражения со словами «симптомы», «проявления», «случаи», «эпизоды», «приступы»

В медицинских текстах часто встречаются выражения вида «*симптомы бронхиальной астмы*», «*эпизоды малопродуктивного кашля*», «*кожные проявления аллергии*» и т.п. Для таких выражений применяются следующие два правила разметки:

- (1) если слова «*симптомы*», «*проявления*» и аналогичные стоят в начале выражения и не имеют дополнительных характеристик, то они не включаются в разметку:

«*симптомы бронхиальной астмы*» — [бронхиальной астмы; *true*]

«*эпизоды малопродуктивного кашля*» — [кашля; *малопродуктивного*]

- (2) если у таких слов есть дополнительные характеристики, то эти слова с их характеристиками включаются в значение признака:

«*кожные проявления аллергии*» — [аллергии; *кожные проявления*]

«*кожные проявления пищевой аллергии*» — [аллергии; *пищевой*],
[аллергии; *кожные проявления*]

«*редкие эпизоды малопродуктивного кашля*» — [кашля; *малопродуктивного*], [кашля; *редкие эпизоды*]

1.3.3. Выражения вида «потеря обоняния», «отсутствие аппетита»

Выражения вида «*отсутствие обоняния*», «*потеря вкуса*», «*снижение аппетита*» и им подобные размечаются в виде булевских признаков со значением *true*: «*жалобы на отсутствие аппетита*» — [отсутствие аппетита; *true*]

1.4. Техническая реализация

Аннотированный текст представляет собой «двумерную» структуру:

- (1) отмеченные сущности типа «признак» и «значение»,
- (2) связи нескольких типов между выделенными сущностями.

Соответственно, чтобы по чистому, неразмеченному тексту получить аналогичную структуру, требуется обучить нейронную сеть извлекать

- (1) сущности заданного типа и
- (2) отношения между ними.

Хотя существуют подходы к обучению нейросети одновременному извлечению сущностей и отношений между ними [12], они носят в основном исследовательский характер. В предлагаемой технической реализации для решения данной задачи используется две различные нейронные сети — одна для извлечения сущностей, вторая — для извлечения отношений. Обе сети реализованы на основе хорошо зарекомендовавших себя для практических целей приложений от компаний Deeppravlov и Facebook, соответственно.

Для обучения нейронных сетей обоих видов использовался один и тот же корпус из примерно 1500 предложений, аннотированных вручную в системе BRAT. В качестве сущностей в предложениях корпуса отмечались сущности

- (1) Feature (признак),
- (2) Value (значение),

где среди признаков специальными атрибутами отмечались (булевские) признаки с отрицанием, а также признаки, представляющие собой анамнез, эпидемиологический статус и общее состояние пациента. Общее количество отмеченных признаков и их значений составило около 5200.

Между сущностями отмечались три вида отношений:

- (1) отношение `has_String` между признаком и значением, отмечающее, что признак имеет символьную характеристику,
- (2) отношение `has_Numerical` между признаком и значением, отмечающее, что признак имеет числовую характеристику,
- (3) отношение `continue` между компонентами признака или между компонентами значения, отмечающее, что второй компонент является продолжением первого.

Данные виды отношений в технической реализации были дополнены отрицательными примерами — парами сущностей, соединенных связью типа «no-relation». Введение отрицательных примеров существенно улучшает результаты обучения нейронной сети извлечению действительных отношений между сущностями. Общее количество положительных примеров (реальных отношений) в корпусе составило около 2400, отрицательных примеров — около 9800. Как обычно, при обучении нейросети извлечению отношений все примеры разбивались на три части — примеры непосредственно для обучения, для проверки во время обучения (валидации) и для тестирования.

Для извлечения сущностей использовалась *предобученная BERT-модель per_rus_bert от компании DeepPavlov^{URL}*. Соответственно, размеченный корпус в BRAT-формате переводился предварительно в стандартный BIO-формат (beginning-inside-outside) для возможности использования этой модели. Точность работы нейронной сети на указанном выше корпусе составила 81,35% (F1-мера).

Аналогично, для извлечения отношений использовалась предобученная BERT-модель SpanBERT [13], предназначенная для предсказания заданных отношений между отдельными (непрерывными) фрагментами текста (спанами). В качестве технической реализации была выбрана *модель от компании Facebook^{URL}*. Далее

- (1) размеченный в BRAT-формате корпус предварительно преобразовывался в *стандартизованный формат TACRED^{URL}* — общепринятый в международном сообществе компьютерной лингвистики формат представления данных для извлечения отношений;
- (2) на основе корпуса в TACRED-формате порождались отрицательные примеры и проходило разбиение всех примеров на обучающую/валидационную/тестовую части (в отношении 75%, 12,5%, 12,5%).

Точность работы обученной нейронной сети по извлечению отношений составила 89,61% (F1-мера).

2. Постановка диагноза

Представляемая система извлечения симптомов и предсказания по ним диагнозов ограничена шестью пульмонологическими заболеваниями (в скобках приведены коды болезней по МКБ-10):

- (1) хроническая обструктивная болезнь легких (J44),
- (2) рак легких и бронхов (C34),
- (3) острый бронхит (J20),
- (4) пневмония (J18),
- (5) бронхит хронический (J41),
- (6) бронхиальная астма (J45),
- (7) COVID-19 (U07.1).

Для обучения классификатора предсказанию диагнозов был создан корпус из 140 текстов (по 20 текстов на каждое заболевание). Каждый текст представляет собой описание конкретного клинического случая с соответствующим диагнозом. Корпус составлен из (неструктурированных) текстов, взятых из материалов, имеющих в открытом доступе в Сети —

медицинских журналов (например, «Инфекционные болезни»), сборников ситуационных (клинических) задач для студентов медицинских вузов и др.

Предварительная обработка этих текстов состояла из двух этапов:

- (1) извлечение диагностических признаков и их значений из каждого текста (с сохранением их в формате json),
- (2) создание обучающего и тестового наборов данных для классификатора болезней.

Классификатор болезней представляет собой простую трехслойную сеть прямого распространения, на которую подаются значения извлеченных из текста описания болезни (медицинской карты пациента) диагностических признаков. Поскольку любая нейронная сеть, в том числе указанного вида, может иметь только фиксированное количество входов, то на основе анализа описаний пульмонологических заболеваний и COVID-19 вручную были отобраны наиболее важные симптомы для заболеваний этого типа. В этот список диагностических признаков вошли:

- (1) **стандартные (выделенные) признаки для данной группы заболеваний** (см. раздел 1.1, принцип II) — «дыхание», «кожные покровы (кожа)», «язык», «слизистые оболочки», «живот»;
- (2) **показатели** — «температура», «пульс», «артериальное давление», «частота дыхания», «сатурация»;
- (3) **булевские признаки** — «чувство нехватки воздуха», «ощущение сдавленности в грудной клетке», «хрипы», «лихорадка», «слабость», «одышка» и др.;
- (4) **признаки со значениями** — «кашель», «рвота», «головная боль», «периферические лимфатические узлы», «задняя стенка глотки» и др.;
- (5) **сведения об анамнезах и общем состоянии пациента** — «общее состояние», «эпидемиологический анамнез», «анамнез жизни».

Некоторые из перечисленных признаков могут иметь значения различных типов. Например, признак «температура» может иметь как числовое значение (39,6°C), так и символьное («повышенная»).

В соответствии с этим общее количество входов-признаков классификатора составило 34 плюс специальный вход «Others», на который подаются признаки с их значениями, которые выделены из исходного текста, но для которых отсутствуют отдельные входы из перечисленных выше. В последнем случае из всех таких признаков с соответствующими им значениями формируется единая строка, которая и подается на вход «Others» классификатора. Хотя эти признаки будут вносить меньший вклад в результирующий диагноз, это позволяет не потерять и учесть всю

диагностическую информацию, которую удалось извлечь из входного документа.

Поскольку входными значениями для классификатора, кроме булевских и числовых, являются символьные выражения (например, для признака «температура» значением может быть выражение «периодически субфебрильная»), то все значения перед подачей на вход классификатора переводятся в числовую форму стандартным способом — с помощью дистрибутивной модели русского языка, а именно, с помощью *BERT-модели rubert-base-cased от компании DeepPavlov*^{URL}. Данная модель переводит каждое выражение в список из 768 вещественных чисел. Для сохранения разумных размеров нейронной сети-классификатора и повышения скорости обучения использовалось только 25 чисел из этого списка. Тем самым, общее количество входов нейронной сети составило $35 \times 25 = 875$. (Эксперименты с использованием большего количества чисел оставлены для дальнейших исследований).

Сама нейронная сеть состоит из 12 нейронов первого слоя, 8 нейронов второго слоя и 7 нейронов (по количеству болезней) выходного слоя. Все слои являются полносвязными (dense-слоями). В качестве функции активации первые два слоя используют кусочно-линейную функцию *relu* (rectified linear unit), а выходной слой — функцию *softmax*, переводящую выходные значения в значения вероятности. Параметры обучения классификатора были стандартными для такого рода сетей, количество итераций обучения — 150.

Кроме перевода символьных значений признаков в числовые, *BERT-модель rubert-base-cased* использовалась для связывания выделенных из текста признаков с конкретными входами классификатора болезнью. Например, у классификатора есть вход «частота дыхания», а в текстах соответствующий признак может формулироваться как «число дыханий», или «частота дыхательных движений», или в виде других (грамматических) форм такого рода выражений. Эта проблема решалась переводом выражений, представляющих признаки, в числовую форму и применением функции, вычисляющей семантическую близость между парами выражений, а именно стандартной функции косинусного сходства. Порог сходства был выбран равным 0.99 — выделенный из текста признак связывался с тем признаком, являющимся входом классификатора, для которого синусное сходство было максимальным и не меньшим 0.99.

Обучение классификатора проводилось на корпусе из 140 текстов с его разбивкой 120 на обучение, 20 — на тестирование. Оценка точности работы классификатора осуществлялась путем кросс-валидации — проводилась 10-кратная случайная разбивка корпуса на обучающую и тестовую части, и подсчитывалось число правильных ответов. Ответ классификатора

считался правильным, если он совпадал с эталонным и степень уверенности (вероятность) его, вычисленная классификатором, превышала 90%. Если хотя бы одно из этих условий нарушалось, ответ классификатора считался неправильным. Точность работы классификатора при кросс-валидационной проверке составила 88,5% (23 случая неправильного диагноза из 200 тестовых). Наиболее трудной для автоматической диагностики оказалась хроническая обструктивная болезнь легких (ХОБЛ) — 7 случаев неправильного диагноза из 23. На следующих местах идут хронический бронхит и пневмония — по 4 случая. На диагностике COVID-19 классификатор не допустил ни одной ошибки.

3. Обзор смежных работ

В данном разделе мы даем обзор наиболее важных работ, посвященных автоматическому прогнозированию диагнозов по анамнезам пациентов, записанных в свободном неструктурированном виде. Специально отмечаются вопросы извлечения диагностических признаков (симптомов) в этих работах. Анализируются методы использования извлеченных симптомов для постановки диагноза. Дается сравнение с аналогичными решениями, принятыми в нашей системе.

В работе [7] заявлено создание алгоритма выявления подозрения на COVID-19. Исходными данными для алгоритма являются симптомы и их значения, извлекаемые из протоколов врачебных осмотров. Количество таких протоколов составило около 11 тыс., однако, установленными диагнозами для них являлись ОРВИ и пневмония, т.е. данные документы не относились непосредственно к COVID-19. Общее количество извлекаемых симптомов составило 14. Они извлекались из текста двумя способами — с помощью методов распознавания именованных сущностей и с помощью специально сконструированных правил. Точность извлечения симптомов составила от 82,6% до 97,43% в зависимости от конкретного симптома. В статье не приводятся сведения о способе извлечения значений таких симптомов как «температура», «частота дыхания», «частота сердечных сокращений» и некоторых других. Сама процедура постановки диагноза базируется на «решающих правилах», построенных на основе мнений врачей-экспертов о правилах взаимодействия симптомов. Результат представляет собой одно из трех значений «уровня подозрения на COVID-19»:

- (1) отсутствует подозрение на COVID-19,
- (2) подозрение на COVID-19,
- (3) вероятен COVID-19.

Количественных оценок точности работы такого алгоритма формирования подозрения на COVID-19 не приведено. Отличие нашей работы от данной состоит в использовании в нашей системе гораздо большего количества анализируемых симптомов, а также в постановке конкретного диагноза заболевания, а не значения «уровня подозрения» на него.

В работе [14] представлена система извлечения симптомов из неструктурированных текстов на английском языке и предсказания наличия COVID-19 по ним. В частности, эта система включает в себя аннотированный корпус из 1472 текстов, содержащих около 30 тыс. упоминаний диагнозов, результатов тестов и симптомов. В силу сходства задачи извлечения симптомов с их значениями с задачей извлечения событий с их аргументами, в данной системе выделение симптомов реализовано как извлечений событий с помощью дообученной Bio+Clinical BERT-модели для решения этой задачи. Точность извлечения симптомов составила 81%, а их значений — от 45% до 78% в зависимости от конкретного значения. Предсказание наличия COVID-19 было реализовано на основе метода случайного леса (random forest classifier) — использовании большого ансамбля решающих деревьев. Данные для обучения такого классификатора включали в себя, кроме симптомов, извлеченных из неструктурированных текстов, 24 структурированных признака. Среди них были

- (1) демографические показатели (пол, возраст),
- (2) жизненно важные показатели (vital signs), такие как температура, частота дыхания и др.,
- (3) результаты лабораторных исследований.

Кроме того, использовалась такая информация как

- (1) сведения, полученные в результате телефонного общения с пациентом (telephone encounters),
- (2) ход амбулаторного лечения (outpatient progress notes),
- (3) обращения к службе скорой помощи (emergency department notes).

Точность предсказания наличия COVID-19 только по структурированным данным составила 61–72%, по симптомам, извлеченным из неструктурированных текстов — 59–67%, по данным обоого типа — 65–78%. Преимущества нашей работы по сравнению с данной системой состоят в том, что наша система имеет возможность предсказывать одно из нескольких заболеваний, обладающих схожими симптомами, что существенно усложняет постановку корректного диагноза. Такое предсказание обеспечивается применением классификатора на основе нейронной сети, которая, в общем случае, является более мощным средством, чем классификатор на основе метода случайного леса.

В работе [4] представлена система принятия диагностических решений на основе данных электронной медицинской карты (ЭМК) пациента, собранных за последние 2 года. Эти данные включают в себя жалобы и анамнез, историю болезни, результаты первичного осмотра, а также лабораторных и инструментальных исследований. Система реализована в научном подразделении SberMedAI Сбербанка России и первоначально носила название «Умный помощник врача TOP-3». Позднее она была преобразована в систему AIDA (AI Diagnostic Assistant) — сервис ИИ для постановки окончательного диагноза на основании данных ЭМК пациента. По информации пресс-службы Сбербанка системы внедрены во всех поликлиниках для взрослых г. Москвы. Общее количество использованных для обучения документов для системы TOP-3 составило первоначально 4 млн., а позднее доведено до 12 млн. Для обучения системы AIDA было использовано 30 млн. записей визитов.

Система TOP-3 выдает результат в виде одного из 265 кодов заболеваний, выбранных из Международной классификации болезней МКБ-10. Система AIDA ограничена 94 наиболее важными диагнозами. Результат работы каждой из систем — выданный диагноз — оценивался врачом. Было использовано два основных вида оценки: по одному и трем выданным возможным диагнозам. В последнем случае, если хотя бы один из трех диагнозов, выданных нейросетью, совпадал с диагнозом врача, то такой ответ засчитывался за правильный. Точность выдачи одного диагноза в системе TOP-3 составила 47,5%, а по трем диагнозам — 68%. (Точность работы диагностических систем на базе ChatGPT также находится в этих пределах — см. об этом ниже). Данные о точности работы системы AIDA отсутствуют.

В качестве модели нейронной сети в обеих системах использовалась незначительно модифицированная стандартная BERT-модель для русского языка. Размер входных текстов ограничивался 128 словами (или, более точно, токенами). Использование такой модели означает, что задача постановки диагноза была сведена к простой задаче классификации текстов на 265 (или 94) категорий. Соответственно, системы такого рода не анализируют медицинские документы на выявление в них таких содержательных единиц как заболевания, симптомы, названия диагностических и лечебных процедур, лекарства и др., а также не определяют отношения между ними. Поэтому данные системы поставят «диагноз» по любому тексту, в том числе и по такому, в котором отсутствуют упоминания симптомов, или вообще не являющемуся медицинским. Кроме того, низкая точность работы таких систем не позволяет их рассматривать в качестве серьезной компоненты систем поддержки клинических решений. Еще одним серьезным их недостатком является отсутствие выдачи

какой-либо информации, объясняющей, почему был поставлен тот или иной диагноз. В нашей системе таблица извлеченных симптомов может служить частичным обоснованием и объяснением выданного диагноза. Кроме того, наша система предоставляет также числовой показатель степени уверенности принятого решения.

В последнее время появилось большое количество работ, описывающих попытки применения больших языковых моделей, в частности, систем типа ChatGPT, для решения задачи постановки медицинских диагнозов. Общий недостаток всех таких систем состоит в том, что их обучение проводится на всей массе данных, имеющихся в Интернете, которые очень часто оказываются неточными и неполными, что критически неприемлемо в задачах медицинской диагностики. Кроме того, такие системы в силу принципов своего построения принципиально не могут рассуждать, проводить простейшие логические умозаключения, которые являются неотъемлемой частью любого процесса постановки диагноза (принятия решения) в любой области. Много работ про тестирование систем на базе ChatGPT в конкретных областях медицины только подтверждают такие выводы.

В работе [10] проверялись возможности постановки диагнозов системой ChatGPT-3.5 в области педиатрии. Для такой оценки были выбраны 100 описаний болезней с подтвержденными диагнозами. Тексты были взяты из архива журнала JAMA Pediatrics и из архива Массачусетского госпиталя общего назначения (Massachusetts General Hospital). Результат, выдаваемый системой ChatGPT, оценивался двумя квалифицированными врачами по трехзначной шкале:

- (1) правильный,
- (2) неправильный,
- (3) не полностью описывает диагноз.

Общее количество ошибочных диагнозов составило 83%, среди них 72% были некорректны, а 11% клинически относящиеся к рассматриваемому случаю, но слишком общие, чтобы их рассматривать в качестве корректного диагноза.

В статье [18] изучались возможности трех систем — GPT-4, Gemini Pro и GPT-4, интегрированной с онтологией фенотипа человека (Human Phenotype Ontology) для постановки диагноза в 61 редком случае педиатрических заболеваний. Система GPT-4 оказалась способной предсказать корректный диагноз в 13,1% случаев, тогда как точность работы систем Gemini Pro и GPT-4 НРО составила только 8,2%.

В работе [19] изучалась точность работы системы GPT-4 в задачах комплексной медицинской диагностики. Исследование проводилось на

70 описаниях историй болезней, взятых из архива журнала New England Journal of Medicine. От системы требовалось два вида ответов:

- (1) наиболее вероятный диагноз,
- (2) дифференциальный диагноз — список возможных диагнозов, ранжированных по степени вероятности.

Во втором случае, если правильный диагноз попадал в данный список, то ответ засчитывался как правильный. Точность работы системы по определению единственного правильного диагноза составила 39% (27 случаев из 70), а в 45 случаях из 70 (64%) правильный ответ присутствовал в списке возможных диагнозов. Средняя длина такого списка для всех случаев составила 9.0.

В статье [20] сообщается о довольно успешных результатах тестирования больших языковых моделей в решении задач диагностики психических расстройств. В этой работе тестировались системы GPT-3.5 Turbo, GPT-4, Аya-101 и Nemo tron. Системы тестировались на 20 клинических случаях, взятых из DSM-5 (Diagnostic and Statistical Manual of mental disorders, fifth edition) — нозологической системе психических расстройств. Ответ систем оценивался по трехзначной шкале:

- (1) правильный,
- (2) частично правильный,
- (3) неправильный.

В результате тестирования, 15 из 20 случаев были диагностированы корректно, по меньшей мере, одной из рассматриваемых систем. Системы GPT-3.5 и Chat GPT-4.0 дали правильный ответ в 13 из 20 случаев, а системы Аya и Nemo tron — в 6–7 случаях из 20. Авторы статьи отмечают, что большое разнообразие психических заболеваний требует создания больших, но, в тоже время, хорошо проработанных датасетов, используемых для обучения моделей.




















4. Выводы и направления дальнейшей работы

В данной работе была представлена система для извлечения симптомов заболеваний из медицинских клинических записей и автоматическому прогнозированию по ним диагноза. Система ограничена 6 пульмонологическими заболеваниями и COVID-19. Подход, реализованный в данной системе, отличается от подхода, принятого в системах TOP-3 и AIDA, разработанных в SberMedAI, и в которых реализована простая классификация текстов. Это позволило существенно повысить точность постановки диагноза, которая составила на указанной группе заболеваний 88.5%. Кроме того, список извлеченных симптомов, на основе которых был поставлен диагноз, можно рассматривать в качестве части его объяснения.



Дальнейшими направлениями совершенствования системы являются:

- (1) повышение точности работы подсистемы извлечения диагностических признаков путем увеличения объема и улучшения качества обучающего множества,
- (2) оценка использования модели RuBioBERT (RuBioRoBERTa) [3] для улучшения связывания выхода подсистемы извлечения симптомов со входом классификатора болезней,
- (3) повышение точности работы классификатора болезней путем более точного подбора симптомов, по которым производится классификация и усовершенствования его внутренней структуры,
- (4) разработка общедоступного Web-варианта системы с использованием из браузера.

Список использованных источников

- [1] Романов Н. А., Сачек М. М. *Системы поддержки принятия клинических решений: современное состояние и проблемы* // Вопросы организации и информатизации здравоохранения.– 2018.– № 3.– С. 18–25.      ↑154
- [2] Osheroff J. A. *Improving medication use and outcomes with clinical decision support: a step-by-step guide.*– HIMSS Publishing.– 2009.– ISBN 9780980069730.– 300 pp. ↑154
- [3] Yalunin A., Nestrov A., Umerenkov D. *RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining.*– 2022.– 5 pp. arXiv     2204.03951  ↑154, 176
- [4] Blinov P., Avetisian M., Kokh V., Umerenkov D., Tuzhilin A. *Predicting clinical diagnosis from patients electronic health records using BERT-based neural networks, Artificial Intelligence in Medicine: 18th Int. Conf. on Art. Intel. In Medicine, AIME 2020 (Minneapolis, MN, USA, August 25-28, 2020), Lecture Notes in Computer Science.*– vol. **12299**, Cham: Springer.– 2020.– ISBN 978-3-030-59136-6.– Pp. 111–121.  ↑154, 155, 173
- [5] Garg R. K., Urs V. L., Agarwal A. A., Chaudhary S. K., Paliwal V., Kar S. K. *Exploring the role ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review* // Health Promotion Perspectives.– 2023.– Vol. **13**.– No. 3.– Pp. 183–191.  ↑154
- [6] Steinkamp J. M., Bala W., Sharma A., Kantrowitz J. J. *Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes* // Journal of Biomedical Informatics.– 2020.– Vol. **102**.– id. 103354.– 9 pp.  ↑154
- [7] Гаврилов Д. В., Кирилкина А. В., Серова Л. М. *Алгоритм формирования подозрения на новую коронавирусную инфекцию на основе анализа симптомов для использования в системах поддержки принятия врачебных решений* // Врач и информационные технологии.– 2020.– № 4.– С. 51–58.       ↑154, 158,

- [8] Umerenkov D., Zubkova G., Nesterov A. *Deciphering diagnosis: how large language models explanations influence clinical decision making*.– 2023.– 11 pp. arXiv:2310.01708 [doi](#) ↑¹⁵⁴
- [9] Tutubalina E., Alimova I., Miftahutdinov Z., Sakhovskiy A., Malykh V., Nikolenko S. *The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews* // *Bioinformatics*.– 2021.– Vol. **37**.– No. 2.– Pp. 243–249. [doi](#) ↑¹⁵⁴
- [10] Barile J., Margolis A., Cason G., Kim R., Kalash S., Tchaconas A., Milanaik R. *Diagnostic accuracy of a large language model in pediatric case studies* // *JAMA Pediatrics*.– 2024.– Vol. **178**.– No. 3.– Pp. 313–315. [doi](#) ↑^{155, 174}
- [11] Сердюк Ю. П., Власова Н. А., Момот С. П. *Система извлечения упоминаний симптомов из текстов на естественном языке с помощью нейронных сетей* // Программные системы: теория и приложения.– 2023.– Т. **14**.– № 1(56).– С. 95–123. [URJ](#) [doi](#) ↑^{158, 159}
- [12] Shang Yu-M., Huang H., Mao X. L. *OneRel: Joint entity and relation extraction with one module in one step* // *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)* (February 22–March 1, 2022, online).– Curran Associates, Inc.– 2022.– ISBN 9781713855743.– Pp. 11285–11293. [URL](#) [doi](#) ↑¹⁶⁷
- [13] Joshi M., Chen D., Liu Y., Weld D. S., Zettlemoyer L., Levy O. *SpanBERT: improving pre-training by representing and predicting spans* // *Transactions of the Association for Computational Linguistics*.– 2020.– Vol. **8**.– Pp. 64–77. [doi](#) ↑¹⁶⁸
- [14] Lybarger K., Ostendorf M., Thompson M., Yetisgen M. *Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework* // *Journal of Biomedical Informatics*.– 2021.– Vol. **117**.– id. 103761.– 13 pp. [doi](#) ↑¹⁷²
- [15] Thirunavukarasu A. J., Ting D. S. J., Elangovan K., Gutierrez L., Tan T. F., Ting D. S. W. *Large language models in medicine* // *Nat. Med.*– 2023.– Vol. **29**.– No. 8.– Pp. 1930–1940. [doi](#) ↑
- [16] Shah N. H., Entwistle D., Pfeffer M. A. *Creation and adoption of large language models in medicine* // *JAMA*.– 2023.– Vol. **330**.– No. 9.– Pp. 866–869. [doi](#) ↑
- [17] Singhal K., Azizi S., Tu T., Mahdavi S. S., Wei J., Chung H. W., Scales N., Tanwani A., Cole-Lewis H., Pfohl S., Payne P., Seneviratne M., Gamble P., Kelly Ch., Babiker A., Schärli N., Chowdhery A., Mansfield Ph., Demner-Fushman D., Agüera y Arcas B., Webster D., Corrado G. S., Matias Y., Chou K., Gottweis J., Tomasev N., Liu Y., Rajkumar A., Barral J., Sementur Ch., Karthikesalingam A., Natarajan V. *Large language models encode clinical knowledge* // *Nature*.– 2023.– Vol. **620**.– No. 7972.– Pp. 172–180. [doi](#) ↑
- [18] Young C. C., Enichen E., Rivera Ch., Auger C. A., Grant N., Rao A., Succi M. D. *Diagnostic accuracy of a custom large language model on rare pediatric disease case reports* // *American Journal of Medical Genetics, Part A*.– 2024.– id. e63878. [doi](#) ↑¹⁷⁴
- [19] Kanjee Z., Crowe B., Rodman A. *Accuracy of generative artificial intelligence model in a complex diagnostic challenge* // *JAMA*.– 2023.– Vol. **330**.– No. 1.– Pp. 78–80. [doi](#) ↑¹⁷⁴

- [20] Gargari G. K., Fatehi F., Mohammadi I., Firouzabadi S. R., Shafiee A., Habibi Gh. *Diagnostic accuracy of large language models in psychiatry* // Asian Journal of Psychiatry.– 2024.– Vol. **100**.– id. 104168.  ↑175
- [21] Nielsen M. *Neural networks and deep learning*, free online book.– 2016.  ↑154

Поступила в редакцию 03.12.2024;
 одобрена после рецензирования 27.12.2024;
 принята к публикации 28.12.2024;
 опубликована онлайн 28.12.2024.

Рекомендовал к публикации

к.т.н. Я. И. Гулиев

Информация об авторах:



Юрий Петрович Сердюк

старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: параллельное программирование, формальные исчисления процессов, системы типов.

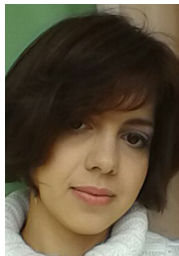
 0000-0003-2916-2102
 e-mail: Yuri@serdyuk.botik.ru



Наталья Александровна Власова

младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: компьютерная лингвистика, автоматическая обработка естественного языка, корпусная лингвистика

 0000-0003-2916-2102
 e-mail: nathalie.vlassova@gmail.com



Седа Рубеновна Момот

инженер Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: компьютерная лингвистика, автоматическая обработка естественного языка

 0000-0003-2916-2102
 e-mail: morlot@mail.ru
morlot@mail.com

Все авторы сделали эквивалентный вклад в подготовку публикации.

Декларация об отсутствии личной заинтересованности: благополучие авторов не зависит от результатов исследования.



Symptoms extraction and automatic diagnosis prediction from medical clinical records

Yuri Serdyuk¹, Natalia Vlasova², Seda Momot³

¹⁻³Ailamazyan Program Systems Institute of RAS, Ves'kovo, Russia

Abstract. The paper introduces a system for symptoms extraction from medical clinical records (texts in natural Russian language) and automatic prediction of a diagnosis in the form of the disease title and its ICD-10 code. The system is designed for a restricted domain of 6 pulmonary diseases (chronic obstructive pulmonary disease, pneumonia, bronchial asthma etc) and COVID-19.

Different neural networks are employed for the symptoms extraction by recognizing certain medical entities and relations between them. A classifier based on a neural network is responsible for the automatic diagnosis. An annotated corpus of sentences is created for the training of the neural networks. The principles and rules of the annotation are described. A corpus of texts is used for the training of the classifier.

Both subsystems were tested, the resulting accuracy estimates are provided. The accuracy of diagnosis in the given domain is 88.5%. We also compare our system with similar works on symptom extraction from texts in various languages, as well as on automatic diagnosis, including systems such as ChatGPT.













(*In Russian*).

Key words and phrases: clinical decision support systems, symptom extraction, automatic diagnosis prediction, BERT models, ChatGPT-based systems.

2020 *Mathematics Subject Classification:* 68T50; 92C50

For citation: Yuri Serdyuk, Natalia Vlasova, Seda Momot. *Symptoms extraction and automatic diagnosis prediction from medical clinical records*. Program Systems: Theory and Applications, 2024, **15**:4(63), pp. 153–181. (*In Russ.*). https://psta.psiras.ru/read/psta2024_4_153-181.pdf

References

- [1] N. A. Romanov, M. M. Sachek. “Clinical decision support systems: current state and problems”, *Voprosy organizatsii i informatizatsii zdravoohraneniya*, 2018, no. 3, pp. 18–25 (in Russian). 
- [2] Osheroff J. A.. *Improving medication use and outcomes with clinical decision support: a step-by-step guide*, HIMSS Publishing, 2009, ISBN 9780980069730, 300 pp.
- [3] A. Yalunin, A. Nestrov, D. Umerenkov. *RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining*, 2022, 5 pp. arXiv:2204.03951 
- [4] P. Blinov, M. Avetisian, V. Kokh, D. Umerenkov, A. Tuzhilin. “Predicting clinical diagnosis from patients electronic health records usin BERT-based neural networks”, Artificial Intelligence in Medicine: 18th Int. Conf. on Art. Intel. In Medicine, AIME 2020 (Minneapolis, MN, USA, August 25-28, 2020), Lecture Notes in Computer Science, vol. **12299**, Springer, Cham, 2020, ISBN 978-3-030-59136-6, pp. 111–121. 
- [5] R. K. Garg, V. L. Urs, A. A. Agarwal, S. K. Chaudhary, V. Paliwal, Kar S. K.. “Exploring the role ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review”, *Health Promotion Perspectives*, **13**:3 (2023), pp. 183–191. 
- [6] J. M. Steinkamp, W. Bala, A. Sharma, Kantrowitz J. J.. “Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes”, *Journal of Biomedical Informatics*, **102** (2020), id. 103354, 9 pp. 
- [7] D. V. Gavrilov, A. V. Kirilkina, Serova L. M.. “Algorithm for forming a suspicion of a new coronavirus infection based on the analysis of symptoms for use in medical decision support systems”, *Vrach i informatzionnye tehnologii*, 2020, no. 4, pp. 51–58 (in Russian).  
- [8] D. Umerenkov, G. Zubkova, A. Nesterov. *Deciphering diagnosis: how large language models explanations influence clinical decision making*, 2023, 11 pp. arXiv:2310.01708 
- [9] E. Tutubalina, I. Alimova, Z. Miftahutdinov, A. Sakhovskiy, V. Malykh, S. Nikolenko. “The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews”, *Bioinformatics*, **37**:2 (2021), pp. 243–249. 
- [10] J. Barile, A. Margolis, G. Cason, R. Kim, S. Kalash, A. Tchaconas, R. Milanaik. “Diagnostic accuracy of a large language model in pediatric case studies”, *JAMA Pediatrics*, **178**:3 (2024), pp. 313–315. 
- [11] Yu. P. Serdyuk, N. A. Vlasova, Momot S. R.. “A system for extracting symptom mentions from texts by means of neural networks”, *Program Systems: Theory and Applications*, **14**:1(56) (2023), pp. 95–123 (in Russian).  

- [12] Yu-M. Shang, H. Huang, X. L. Mao. “OneRel: Joint entity and relation extraction with one module in one step”, *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)* (February 22–March 1, 2022, online), Curran Associates, Inc., 2022, ISBN 9781713855743, pp. 11285–11293. [URL](#) [doi](#)
- [13] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy. “SpanBERT: improving pre-training by representing and predicting spans”, *Transactions of the Association for Computational Linguistics*, **8** (2020), pp. 64–77. [doi](#)
- [14] K. Lybarger, M. Ostendorf, M. Thompson, M. Yetisgen. “Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework”, *Journal of Biomedical Informatics*, **117** (2021), id. 103761, 13 pp. [doi](#)
- [15] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, Ting D. S. W.. “Large language models in medicine”, *Nat. Med.*, **29**:8 (2023), pp. 1930–1940. [doi](#)
- [16] N. H. Shah, D. Entwistle, M. A. Pfeffer. “Creation and adoption of large language models in medicine”, *JAMA*, **330**:9 (2023), pp. 866–869. [doi](#)
- [17] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, Ch. Kelly, A. Babiker, N. Schärli, A. Chowdhery, Ph. Mansfield, D. Demner-Fushman, Agüera y Arcas B., D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, Ch. Semturs, A. Karthikesalingam, V. Natarajan. “Large language models encode clinical knowledge”, *Nature*, **620**:7972 (2023), pp. 172–180. [doi](#)
- [18] C. C. Young, E. Enichen, Ch. Rivera, C. A. Auger, N. Grant, A. Rao, M. D. Succi. “Diagnostic accuracy of a custom large language model on rare pediatric disease case reports”, *American Journal of Medical Genetics, Part A*, 2024, id. e63878. [doi](#)
- [19] Z. Kanjee, B. Crowe, A. Rodman. “Accuracy of generative artificial intelligence model in a complex diagnostic challenge”, *JAMA*, **330**:1 (2023), pp. 78–80. [doi](#)
- [20] G. K. Gargari, F. Fatehi, I. Mohammadi, S. R. Firouzabadi, A. Shafiee, Gh. Habibi. “Diagnostic accuracy of large language models in psychiatry”, *Asian Journal of Psychiatry*, **100** (2024), id. 104168. [doi](#)
- [21] M. Nielsen. *Neural networks and deep learning*, free online book, 2016. [URL](#)