


УДК 81'322+61

 10.25209/2079-3316-2023-14-1-95-123

## Система извлечения упоминаний симптомов из текстов на естественном языке с помощью нейронных сетей

Юрий Петрович Сердюк<sup>1</sup>, Наталья Александровна Власова<sup>2✉</sup>,  
Седа Рубеновна Момот<sup>3</sup>

Институт программных систем им. А. К. Айламазяна РАН, Вельсково, Россия

<sup>2✉</sup> [nathalie.vlassova@gmail.com](mailto:nathalie.vlassova@gmail.com)

**Аннотация.** В статье представлена система для извлечения упоминаний симптомов из медицинских текстов на естественном (русском) языке. Система осуществляет нахождение симптомов в тексте, их нормализацию (приведение к стандартной форме) и отождествление – отнесение найденного симптома к группе однотипных симптомов. Каждый этап обработки реализуется с помощью отдельной нейронной сети. Состав извлекаемых симптомов ограничен тремя видами заболеваний – аллергические и пульмонологические заболевания, а также коронавирусная инфекция (COVID-19). Представлен и описан аннотированный корпус предложений, использованный для обучения нейросети нахождению упоминаний симптомов, относящихся к этим трем заболеваниям. При разметке корпуса был использован простой XML-подобный язык. Для представления предложений, непосредственно поступающих на вход нейросети, предложен расширенный ВЮ-формат разметки. Для каждого этапа приведены оценки точности (для первого этапа точность оценивалась при строгом и гибком тестировании). Описаны подходы и реализация приведения к стандартной форме и отождествления упоминаний симптомов. Даны сравнения с аналогичными работами по извлечению симптомов из медицинских текстов на разных языках, а также показано место данной системы в системах поддержки принятия клинических решений.

**Ключевые слова и фразы:** автоматическая обработка языка, нейронные сети, автоматическое извлечение информации, аннотированный корпус, упоминания симптомов, BERT-модели, Covid-19

**Для цитирования:** Сердюк Ю.П., Власова Н.А., Момот С.Р. 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. [https://psta.psir.ru/read/psta2023\\_1\\_95-123.pdf](https://psta.psir.ru/read/psta2023_1_95-123.pdf)

## Введение

Медицинские системы поддержки принятия решений являются, с одной стороны, важной компонентой современного подхода к оказанию врачебной помощи, а с другой – высокотехнологичной областью, впитывающей в себя последние разработки в области искусственного интеллекта (ИИ) [1, 2]. Эффективность их применения во многом зависит от состава и качества информации, которая доступна врачам при постановке диагноза и проведении лечения. В свою очередь, эта информация распределена между ее источниками в структурированной форме – базах данных, и в неструктурированной – в виде разнообразных медицинских документов на естественном языке (ЕЯ), в состав которых входят истории болезней, анамнезы, эпикризы, отчеты о проведении клинических мероприятий (анализов, операций, обследований). Соответственно, извлечение и структурирование необходимых сведений из текстов на ЕЯ является критически важной задачей, от качества решения которой во многом зависит эффективность работы всей системы поддержки принятия клинических решений.

Из медицинских ЕЯ-текстов в общем случае требуется извлечь сущности разных типов – названия заболеваний и нарушений функционирования человеческого организма, симптомов болезней, медицинских процедур, лекарственных препаратов, упоминания частей тела и другие. Извлечение каждой такой сущности из текста на ЕЯ является достаточно сложной задачей, в состав которой включается не только сам этап извлечения, но и этапы нормализации (приведения к стандартной форме) и отнесения извлеченной сущности к предопределенному словарю (тезаурусу) или категории.

В данной работе мы предлагаем комплексный подход к извлечению упоминаний симптомов болезней из текстов на ЕЯ, включающий в себя все три отмеченных выше этапа – само извлечение, нормализацию и отождествление. В общем случае возможны два подхода к извлечению симптомов:

- (1) извлечение упоминаний симптомов и
- (2) извлечение (названий) симптомов с их значениями.

В первом случае под *упоминанием симптома* понимается некоторая непрерывная (без разрывов и пропусков) часть текста, которая используется врачами-практиками в качестве устойчивого словосочетания для описания некоторого симптома:

Мальчик поступил в отделение с жалобами на **затруднение носового дыхания**.

Во втором, более общем и сложном случае, выражение, представляющее собой описание симптома, рассматривается в структурированном виде, а именно в виде пары *⟨название симптома, значение симптома⟩*:

*Температура тела поднимается до 38°C, либо частота дыхательных движений доходит до 21-22 в минуту.*

Часто, как в приведенном выше примере, название симптома и его значение могут находиться в предложении на некотором «расстоянии» друг от друга, т.е. не составлять непрерывное выражение.

Соответственно, в первом случае при аннотировании упоминаний симптомов в тексте достаточно отметить начало и конец размечаемого выражения, что делает задачу извлечения упоминания симптомов схожей с задачей извлечения именованных сущностей (NER – named entity recognition), таких как «*Высшая школа экономики*» или «*Национальный Совет молодёжных и детских объединений России*».

Во втором случае требуется применение «двумерного» аннотирования – разметки по отдельности названий симптомов и их значений вместе с установлением отношений между ними. При таком подходе задача извлечения симптомов с их значениями становится схожей с задачей извлечения событий с их аргументами [3] или задачей извлечения семантических отношений [4].

В данной работе мы представляем решение первой задачи – извлечения упоминаний симптомов. Состав извлекаемых симптомов ограничен в нашем случае тремя видами заболеваний:

- аллергические заболевания,
- пульмонологические заболевания,
- коронавирусная инфекция (COVID-19).

Основу корпуса предложений с упоминаниями симптомов аллергических и пульмонологических заболеваний составил *корпус клинических текстов* <sup>(URL)</sup>, подготовленный Институтом системного анализа ФИЦ ИУ РАН и ФГНБНУ «Научный центр здоровья детей» (НЦЗД) в 2015 г. Корпус предложений с упоминаниями симптомов коронавирусной инфекции был подготовлен нами самостоятельно на основе медицинских документов и текстов, имеющихся в Интернете.

Для аннотирования упоминаний симптомов мы применяем простой XML-подобный язык. При подаче непосредственно на вход нейросети он переводится в расширенный BIO-формат (BIO – beginning, inside, outside) по сравнению с его стандартным вариантом, применяемым в NER-задачах. Для более формальной трактовки понятий «симптом» и

«упоминание симптома» нами были введены понятия «симптом-отклонение» и «симптом-характеристика». Кроме того, мы проводим различия между «простым симптомом» и «перечислением симптомов». Более подробные сведения о корпусе предложений для извлечения упоминаний симптомов, принципах и правилах их аннотирования приведены в разделе 1.

Для практического решения задачи извлечения упоминаний симптомов как задачи распознавания именованных сущностей мы применили систему *DeepPavlov*<sup>URL</sup> – библиотеку для анализа текстов на ЕЯ и создания виртуальных помощников (чат-ботов), разработанную в Лаборатории нейронных сетей и глубокого обучения МФТИ. Данная библиотека содержит несколько типов моделей для решения NER-задач, начиная от стандартных моделей на основе рекуррентных нейронных сетей, а также BERT-моделей и заканчивая гибридными моделями, объединяющими в себе несколько нейронных сетей различного типа. Нами была выбрана гибридная модель в силу ее небольшого размера и достаточно хороших показателей точности извлечения [5]. В частности, был разработан и использован собственный формат тегов разметки упоминаний симптомов – расширенный BIO-формат. Подробности практической реализации этого подхода даются в разделах 1.3 и 1.4.

После извлечения выражения, являющегося упоминанием симптома, оно приводится в нормальную форму. Так, например, из предложения

*Ребенок поступил в отделение впервые с жалобами на эпизоды  
малопродуктивного кашля*

должно быть извлечено выражение *малопродуктивного кашля* с дальнейшим переводом его в нормальную форму *малопродуктивный кашель*.

До недавнего времени для решения этой задачи применялись методы на основе морфо-синтаксического анализа, в которых каждое слово выражения преобразовывалось в корректную форму с учетом синтаксических связей этого слова с другими словами выражения. С точки зрения реализации эти методы основывались на использовании множества правил, описывающих соответствующие преобразования. Данный подход обладает тем недостатком, что для качественного выполнения нормализации требуется очень большое количество правил, составление которых является трудоемкой задачей. Кроме того, увеличение количества правил неизбежно приводит к появлению противоречий между ними.

В последнее время для решения задач нормализации выражений, например, именованных сущностей и им подобных, начали применяться методы на основе нейронных сетей [6]. В этом случае нейронная сеть обучается на предложениях с выделенными выражениями, для которых

подготовлены их стандартные представления. В частности, для более полного учета семантических свойств отдельных слов используются BERT-модели. Для решения данной задачи в настоящей работе мы воспользовались системой, разработанной Д. Анастасьевым [7]. Структура этой системы, состав и особенности корпуса для нормализации выражений, а также детали практической реализации и полученные результаты изложены в разделе 2.

Задачей нормализации выражений проблема извлечения упоминаний симптомов не исчерпывается. Конкретный симптом как семантическое понятие может иметь много форм своего выражения. Например, выражения *затруднение дыхания*, *затрудненное дыхание*, *нарушение функции дыхания*, *приступы затруднения дыхания*, *дыхательная недостаточность 2 степени*, *одышка* обозначают схожие явления. Привязывание каждого из такого типа выражений к некоторому стандартизованному (эталонному) выражению иногда называют задачей нормализации понятий (concept normalization) или связыванием сущностей (entity linking).

В некоторых случаях в качестве ресурса эталонных названий тех или иных медицинских сущностей (заболеваний, симптомов, лекарств, процедур и др.) используются разного вида тезаурусы и онтологии. Одним из главных примеров здесь является онтология UMLS (Unified Medical Language System) [9]. Составной частью онтологии UMLS является тезаурус предметных медицинских рубрик (Medical Subject Headings) [10], который имеет эквивалент на русском языке MSHRUS, поддерживаемый Центральной Научной Медицинской Библиотекой (ЦНМБ).

С другой стороны, в последнее время компанией Соцмедика<sup>URL</sup> разрабатывается объединенная база медицинских знаний UMKB (Unified Medical Knowledge Base) [11], составной частью которой является классификатор медицинских терминов.

Тем не менее имеются большие проблемы в совместном использовании различных онтологий и тезаурусов на русском языке. Поэтому в нашем случае мы разработали собственную систему базовых (эталонных) терминов для упоминаний симптомов аллергических и пульмонологических заболеваний, а также для наиболее важных симптомов коронавирусной инфекции. Механизм привязывания извлеченных упоминаний симптомов к выбранным эталонным терминам изложен в разделе 3.

В разделе 4 дан обзор современных работ по извлечению упоминаний симптомов, а также симптомов в виде пар *⟨название симптома, значение симптома⟩* и проведено их сравнение с нашей работой.

Раздел 5 посвящен кратким выводам и направлениям дальнейшей работы.

## 1. Извлечение упоминаний симптомов

### 1.1. Принципы и правила аннотации упоминаний симптомов

#### 1.1.1. Общие принципы выделения симптомов

Одно из классических определений понятия «симптом» описывает его как «признак болезни, качественно новый, не свойственный здоровому организму феномен» (Большая медицинская энциклопедия, 1989 г., 3-е изд.). Однако в реальной медицинской практике понятие «симптом» трактуется более широко – от описаний всевозможных отклонений (*сухой кашель*) до названий болезней, которые могут являться симптомами других болезней (*ринит, конъюнктивит*), выражений, включающих числовые данные (*частота дыхательных движений 21-22 в минуту*), характеристик отдельных органов и их функционирования (*сердечные тоны ясные*) и др.

Ввиду принципиальной невозможности строгим формальным образом определить понятие симптома (и, соответственно, его упоминания), в качестве основного принципа аннотирования предлагается выделение из текста в качестве упоминаний симптомов таких выражений, которые

- (1) используются врачами-практиками в качестве устойчивых словосочетаний,
- (2) обозначают либо «симптом-отклонение», либо «симптом-характеристику».

Далее в тексте статьи для легкости восприятия мы часто употребляем термин «симптом» в значении «упоминание симптома».

«*Симптомы-отклонения*» – симптомы, само название которых говорит о некотором отклонении от нормы (*малопродуктивный кашель, хрипы, дизурия, менингеальные знаки*):

- (а) жалобы на ухудшение общего состояния: *нарушение сна, повышенная утомляемость, снижение работоспособности*;
- (б) симптомы, непосредственно наблюдаемые или ощущаемые самим пациентом: *сухой кашель, одышка, боль в пояснице, высыпания*;
- (в) нарушения, фиксируемые врачом при осмотре/обследовании: *хрипы в легких, менингеальные знаки, нарушения бронхиальной проходимости*.

«*Симптомы-характеристики*» – это выражения, в которых присутствует указание на некоторую часть тела или физиологический процесс с соответствующими им характеристиками (*живот мягкий, тоны сердца ритмичные*). Необходимо отметить, что общепринятая врачебная практика состоит в использовании таких выражений в качестве упоминаний симптомов независимо от того, обозначают ли они некоторое отклонение от нормы (отрицательная характеристика), или же нормальное функционирование организма или его подсистемы (положительная характеристика):

- (1) отклонение от нормы (отрицательная характеристика): *носовое дыхание несколько затруднено, воронкообразная деформация грудной клетки;*
- (2) нормальное функционирование органа или системы (положительная характеристика): *кожные покровы бледно-розовые, зев спокоен, сердечные тоны ясные.*

Упоминания симптомов обоих типов могут сопровождаться указанием на:

- (1) характер проявления симптома (*малопродуктивный кашель, схваткообразные боли*);
- (2) интенсивность проявления симптома (*умеренные боли*);
- (3) часть тела, к которой относится симптом (*боли в суставах*);
- (4) временные рамки и условия проявления (*ночное недержание мочи, аллергический ринит в весенне-летний период, одышка при физической нагрузке*).

### 1.1.2. Правила аннотирования симптомов

Для аннотирования упоминаний симптомов был использован простой XML-подобный язык разметки:

*Мальчик поступил в отделение с жалобами на <symp> затруднение носового дыхания </symp> в весенний и осенний периоды.*

Упоминания симптомов в тексте делятся на два вида:

- (а) простые (единичные) симптомы,
- (б) перечисления симптомов.

Простые симптомы выделяются с помощью тегов <symp> и </symp>:

*У ребенка 2 лет на фоне ОРВИ внезапно ночью возник <symp> лающий кашель </symp>, <symp> охриплость </symp>.*

Перечисление симптомов – это выражения с упоминаниями нескольких симптомов, которые должны быть извлечены по отдельности. Так, из выражения *Сердечные тоны ясные, ритмичные* должны быть выделены простые симптомы

- *Сердечные тоны ясные,*
- *Сердечные тоны ритмичные*

Соответственно, главная цель разметки таких выражений – отметить составные части отдельных симптомов.

Перечисления симптомов обычно состоят из основной части и подчиненной части. При этом возможны два варианта:

- (1) одна основная и несколько подчиненных, и
- (2) несколько основных частей и одна подчиненная.

(Случаи, когда выражение состоит из нескольких основных и нескольких подчиненных частей, возможны, но очень редки). В типичном случае основная часть содержит названия частей тела или физиологических процессов, а подчиненная – некоторые характеристики этих частей или процессов (эти характеристики мы называем аргументами).

**Вариант 1.** Перечисление симптомов с одной основной и несколькими подчиненными частями заключается в теги `<symp-args>` и `</symp-args>`, внутри которых отдельные аргументы отмечаются тегами `<arg>` и `</arg>`:

```
<symp-args> Сердечные тоны <arg> ясные </arg>, <arg> ритмичные </arg>
</symp-args>.
```

```
<symp-args> Живот <arg> мягкий </arg>, <arg> при пальпации безболезненный
</arg> </symp-args>.
```

**Вариант 2.** Перечисление симптомов с несколькими основными частями и одной подчиненной заключается в теги `<symp-args>` и `</symp-args>`, внутри которых отдельные основные части отмечаются тегами `<symp>` и `</symp>`, а подчиненная – тегами `<arg>` и `</arg>`.

```
<symp-args> <symp> Стул </symp>, <symp> диурез </symp> <arg> не нарушены
</arg> </symp-args>.
```

```
У пациентов возникают <symp-args> <symp> сухость </symp> и <symp>
неприятные ощущения </symp> <arg> в носоглотке </arg> </symp-args>.
```

### 1.1.3. Специальные случаи

Как уже было отмечено выше, в реальной медицинской практике понятие симптома трактуется очень широко. В данном разделе рассматриваются несколько сложных специальных случаев упоминания симптомов и правила их разметки, общепринятые у врачей-аннотаторов.

#### 1. Заболевание как симптом.

Встречаются следующие случаи использования названия заболеваний в качестве симптомов:

- (а) заболевание как симптом других заболеваний:

```
В последнее время все больше внимания привлекает <symp>
конъюнктивит </symp> как новый симптом коронавируса.
```

- (б) название заболевания, которому предшествует слово *симптомы*:



*Мальчик поступает в клинику впервые с жалобами на <symp> симптомы бронхиальной обструкции </symp>.*

- (6) название заболевания, которому предшествуют слова, указывающие на эпизодический характер его проявления, например, *эпизоды, проявления, случаи, приступы:*

*В течение нескольких лет отмечались <symp> проявления бронхиальной астмы </symp>.*

*<symp> Эпизоды бронхиальной обструкции </symp> отмечались 3-4 раза в год.*

С другой стороны, если аналогичные слова стоят перед упоминанием симптома-отклонения, то они не включаются в размечаемое выражение:

*эпизоды <symp> малопродуктивного кашля </symp>.*

- (2) заболевания, сопровождающееся указанием на регулярные рецидивы: *частые, рецидивирующие:*

*В дальнейшем <symp> частые ОРИ </symp>, <symp> рецидивирующие аденоидиты </symp>.*

К этой же группе относится и упоминание симптома вида *частые смены настроения*, иногда встречающееся в медицинских документах. С другой стороны, не отмечаются в качестве упоминаний симптомов выражения вида *часто болеющие дети, часто болеющий ребенок* и т. п., поскольку они самостоятельно классифицируются как заболевания в МКБ-10 (D84.9 «Иммунодефицит неуточненный»).

## 2. Симптомы с уточняющими характеристиками.

В медицинских текстах симптомы часто сопровождаются дополнительными характеристиками, такими как интенсивность их проявления, локализация, условия и временные особенности проявления. Все или некоторые такие характеристики могут быть включены в размечаемое выражение, если они составляют вместе с симптомом непрерывный фрагмент текста. Так, например, допускаются следующие варианты разметки одного и того же выражения:

*С января 2030 года беспокоят <symp> высыпания </symp> на коже ягодиц.*

*С января 2030 года беспокоят <symp> высыпания на коже </symp> ягодиц.*

*С января 2030 года беспокоят <symp> высыпания на коже ягодиц </symp>.*

Такой подход к разметке мы называем «гибким подходом», и он отображает реальную практику разметки аналогичных выраже-

ний врачами-аннотаторами. Приведем несколько дополнительных примеров разметки симптомов с уточняющими характеристиками:

*Девочка 8 лет, почувствовала <symp> боль в горле при глотании </symp>.*

*И больной астмой грудной ребенок, и его сверстник без признаков аллергии могут дать <symp> эпизод обструкции на фоне ОРВИ </symp>.*

## 1.2. Корпус предложений для обучения и тестирования

Для обучения и тестирования нейронной сети использовался корпус предложений из медицинских текстов объемом около 1,1 тыс предложений (неповторяющихся). Тематика текстов-источников – аллергические, пульмонологические заболевания и COVID-19. Все предложения аннотированы вручную с помощью специально разработанной системы тегов.

### 1.2.1. Корпус предложений с симптомами аллергических и пульмонологических заболеваний

В качестве основы корпуса предложений с симптомами аллергических и пульмонологических заболеваний использован *корпус медицинских текстов* <sup>URL</sup>, подготовленный ИСА РАН и Научным центром здоровья детей (НЦЗД) [4]. Этот корпус содержит анонимизированные медицинские истории более 60 пациентов НЦЗД (всего 112 текстов). Тексты корпуса включают в себя выписные эпикризы из историй болезней, заключения различного вида обследований, а также назначения и рекомендации различных врачей. Для использования в научных целях данный корпус доступен по запросу.

Корпус размечен практикующими врачами и включает аннотации таких сущностей, как заболевание (Disease), симптом (Symptom), тяжесть заболевания (Severity), течение болезни (Course), назначения (Treatment), медикаменты (Drug), части тела (Body location). Для отдельных видов сущностей установлены отношения между ними.

В силу сложности выделения перечисленных сущностей и отношений между ними разметка симптомов в данном корпусе иногда является непоследовательной и противоречивой. В частности, бывает трудно разбить выражение на сам симптом и место (часть тела) его проявления или выделить в сложном выражении отдельные симптомы. Так, например, в качестве симптома (Symptom) в данном корпусе встречаются такие выражения, как *мягкий; безболезненный; живот мягкий, безболезненный; боли; головные боли; боли в животе и головные боли; боли и потеря чувствительности.*

В нашем случае из этого корпуса было взято около 500 неповторяющихся предложений. Для некоторых симптомов, которые были представлены недостаточным образом в ранее отобранных предложениях,

было добавлено еще около 150 предложений из медицинских текстов (научные статьи, истории болезни, ситуативные задачи из учебников медвузов и т. п.).

### 1.2.2. Корпус предложений с симптомами коронавирусной инфекции

Вторая часть корпуса для обучения извлечению упоминаний симптомов подготовлена нами самостоятельно и состоит из множества предложений, относящихся к новой коронавирусной инфекции. Общее количество предложений в этом подкорпусе – около 500. В каждом предложении вручную размечены симптомы COVID-19 согласно разработанным принципам и правилам аннотации упоминаний симптомов (см. раздел 1.1.1).

Корпус создан на основе представленных в Интернете медицинских текстов, в которых содержатся описания симптомов и проявлений новой коронавирусной инфекции, например, таких как [13]. В некоторых работах [18, 19], упоминается от 10 до 60 разнотипных симптомов данной инфекции, которые извлекались из текстов. Нами было выделено 15 наиболее важных симптомов, характеризующих данное заболевание. В их состав вошли:

- |                    |                        |                        |
|--------------------|------------------------|------------------------|
| (1) температура;   | (7) потеря обоняния и  | (11) насморк;          |
| (2) кашель;        | вкуса;                 | (12) затруднение дыха- |
| (3) боль в груди;  | (8) диарея;            | ния;                   |
| (4) одышка;        | (9) головная боль;     | (13) кожные высыпания; |
| (5) слабость;      | (10) учащенное сердце- | (14) рвота;            |
| (6) боль в мышцах; | биение;                | (15) поражения глаз.   |

Для каждого из 15 симптомов было подобрано и размечено не менее, чем по 30 предложений с различными формами упоминания этих симптомов. Так, например, для симптома «кашель» были размечены такие выражения, как *кашель; частый кашель с небольшим количеством мокроты; малопродуктивный кашель; кашель с мокротой; постоянное покашливание; легкое покашливание; непрерывный кашель; кашель становится сухим и стойким; сухой, часто надсадный кашель; сухой ковидный кашель* и т. п.

Упоминания симптомов коронавирусной инфекции являются более разнообразными, а по своей структуре – более сложными, чем упоминания симптомов аллергических и пульмонологических заболеваний. В частности, они могут включать в себя выражения с числовыми данными:

*Большая 36 лет, госпитализирована с жалобами на <symp> повышение температуры до 39,8° С </symp>, <symp> слабость </symp>, <symp> утомляемость </symp>, <symp> кашель с трудноотделяемой мокротой </symp>, <symp> чувство нехватки воздуха </symp>.*

### 1.3. Расширенный ВГО-формат разметки упоминаний симптомов

Прежде чем текст с XML-подобной разметкой будет подан на вход нейронной сети для обучения, он переводится в так называемый «расширенный ВГО-формат». Расширение обычного ВГО-формата необходимо для представления перечисления симптомов (см. раздел 1.1.2). Для представления упоминаний простых симптомов используется стандартная ВГО-разметка:

```
У ребенка <symp> проявления atopического дерматита </symp> .
O      O                B-SYM      I-SYM      I-SYM      O
```

Здесь метки **B-SYM** и **I-SYM** являются аналогами меток **B** и **I** стандартного ВГО-формата.

Таким образом, в процессе обучения на вход нейронной сети последовательно подаются отдельные слова (а точнее их векторные представления – эмбединги, см. раздел 3) с одновременным предъявлением ей правильных выходов – меток ВГО-формата.

Для представления перечислений симптомов используются дополнительные виды меток. Так, например, выражения с аргументами представляются следующим образом:

```
<symp-args> Живот <arg> мягкий </arg> , <arg> безболезненный </arg>
              B-SA          B-ARG      O          B-ARG
</symp-args>
              E-SA
```

Здесь метки **B-SA** (*begin-symptom with arguments*) и **E-SA** (*end-symptom with arguments*) обрамляют всё выражение с перечислением симптомов, а метки **B-ARG** (*begin-argument*), а также **I-ARG** (*inside-argument*) отмечают отдельные аргументы.

Аналогично выражение с перечислением симптомов второго типа переводится в расширенный ВГО-формат следующим образом:

```
<symptom-list> <symp> Стул </symp> , <symp> диурез </symp> <arg> не нарушены
              B-SL      O          B-SL          B-ARG I-ARG
</arg> </symptom-list>
              O
```

Здесь метки **B-SL** отмечают начало отдельного симптома в списке симптомов. Отдельный симптом может состоять из нескольких слов, а потому для разметки продолжения симптома используются метки **I-SYM**.

Соответственно, после обучения на этапе работы нейросети она пытается разметить входное предложение с помощью меток расширенного ВГО-формата. Используя эти метки, система из представления

```
Живот мягкий , безболезненный .
B-SA B-ARG O B-ARG E-SA
```

восстановит перечисление симптомов в виде *живот мягкий, живот безболезненный*.

Аналогично по расширенной BIO-разметке

*Стул , диурез не нарушены .*  
*B-SL O B-SL B-ARG I-ARG O*

будет восстановлено перечисление симптомов *стул не нарушен, диурез не нарушен*.

При этом будет выполнена попытка требуемого в общем случае синтаксического согласования отдельных частей выражения. Непосредственно выделенное выражение *стул не нарушены* будет переведено в выражение *стул не нарушен*.

Если нейронная сеть не формирует для некоторого предложения корректную BIO-разметку, то считается, что система не нашла в этом предложении упоминаний симптомов.

Машинная реализация расширенного BIO-формата базируется на возможности использования произвольных меток для разметки именованных сущностей в системе *DeepPavlov*<sup>URL</sup>.

#### 1.4. Практическая реализация и результаты

Для практического решения задачи извлечения упоминаний симптомов была использована система *DeepPavlov*<sup>URL</sup> – библиотека и набор моделей для анализа ЕЯ-текстов. В частности, данная система содержит подсистему для распознавания именованных сущностей, которая в прямом виде может быть применена для извлечения упоминаний симптомов. В этой подсистеме имеется несколько типов моделей для решения NER-задач. Нами выбрана гибридная модель, объединяющая в себе несколько нейронных сетей различного типа, в силу ее небольшого объема и достаточно хороших показателей точности распознавания [5]. (В последних версиях системы *DeepPavlov*<sup>URL</sup> предлагается только один тип моделей – BERT-модели).

Данная гибридная модель состоит из

- (1) сверточной (convolutional) нейронной сети, ответственной за распознавание морфологической структуры слов, в частности, за распознавание префиксов и суффиксов;
- (2) двунаправленной LSTM-сети (bidirectional long-short term memory – bi-LSTM) – специального вида рекуррентной сети, ответственной за распознавание слов, написанных с заглавной буквы (что является спецификой общей задачи распознавания именованных сущностей, но иногда применимо и для распознавания симптомов типа «отеки Квинке»);

- (3) еще одной bi-LSTM-сети, ответственной за распознавание правого и левого контекстов слова;
- (4) распознавателя на базе CRF-слоя (conditional random field) (вместо заключительного стандартного softmax-слоя), ответственного за более точное распознавание порядка следования тегов разметки.

Для числового векторного представления слов использовалась предобученная модель, построенная на основе корпуса новостных текстов с русскоязычного сайта *lenta.ru*. Длина вектора составляла 100 чисел. Обучающее множество содержало около 1000 предложений. Размер словаря различных словоформ упоминаний симптомов в этом множестве также составил около 1000 выражений. Тестовое множество состояло из 90 предложений со 160 упоминаниями симптомов. Это же множество использовалось как валидационное на этапе обучения нейронной сети.

Было проведено 2 вида тестирования:

- (1) «строгое», когда нейросеть должна была в точности выделить заданный список симптомов;
- (2) «гибкое» (см. раздел 1.1.3 про гибкий подход при разметке симптомов с уточнениями), когда возможны несколько вариантов выделения упоминания симптома из выражения, и выделение одного из них засчитывается за правильный ответ.

Точность распознавания упоминаний симптомов при строгом тестировании составила 81–82%, а при гибком – 85–86%.

## 2. Нормализация выражений

После этапа извлечения выражений, представляющих собой упоминания симптомов, выполняется нормализация этих выражений. Необходимость ее вызвана тем, что в общем случае извлеченные упоминания симптомов не находятся в нормализованной форме: *малопродуктивным кашлем; одышку, возникшую после стресса; обструктивного ларинготрахеита; хрипов* и т. п.

Способ решения задачи нормализации такого вида выражений с помощью нейронных сетей состоит в том, что на этапе обучения нейросети на вход подаются предложения с выделенными выражениями одновременно с примерами стандартных представлений этих выражений. Таким образом, приведенные выше выражения должны быть в итоге переведены нейронной сетью в выражения *малопродуктивный кашель; одышка, возникшая после стресса; обструктивный ларинготрахеит; хрипы*.

Формально нормализация выражения состоит в переводе главного слова выражения (корня синтаксической структуры) в именительный

падеж с соответствующим согласованием синтаксических связей данного слова с другими словами выражения. Следует отметить, что задача нормализации медицинских выражений отличается от классической задачи лемматизации – перевода слова еще и в единственное число, так как в медицинских документах некоторые выражения стандартно используются только во множественном числе: *хрипы*, *менингеальные знаки* и др. Таким образом, для нормализации такого рода выражений необходимо иметь в обучающем множестве соответствующие примеры с ними.

Возможны два подхода к решению задачи нормализации выражений, представляющих собой упоминания симптомов, с помощью нейронных сетей:

- (1) обучение нейронной сети на специальном датасете всевозможных форм упоминаний симптомов (из заданной предметной области) с соответствующими им стандартными представлениями;
- (2) обучение нейронной сети на датасете с выражениями общего вида с дальнейшим применением ее к нормализации выражений, представляющих симптомы.

При первом подходе требуется подготовка собственного, достаточно большого обучающего множества предложений с упоминаниями симптомов из заданной предметной области, что является довольно трудоемкой задачей. Нами выбран второй подход, поскольку уже существуют аннотированные русскоязычные корпуса, предназначенные для обучения нейронных сетей нормализации выражений общего вида, а также несколько практических реализаций систем, решающих эту задачу.

В нашем случае мы взяли систему, разработанную Д. Анастасьевым [7], которая, в свою очередь, базируется на морфо-синтаксическом анализаторе (а точнее, лемматизаторе из него) из работы [8]. Данная система показала один из наиболее высоких результатов в рамках соревнования RuNormAS-2021 «A shared task on Russian normalization of annotated spans», состоявшегося в рамках конференции «Диалог-2021» [20]. В качестве обучающего множества мы взяли часть корпуса, предоставленного на этом соревновании. Общее количество выражений (как общего вида, так и именованных сущностей) составило около 20 000. Для учета специфики некоторых медицинских терминов из области аллергических и пульмонологических заболеваний, а также коронавирусной инфекции, нами было подготовлено и добавлено в это множество еще около 100 предложений.

Базовой частью системы Д. Анастасьева является лемматизатор, который порождает все возможные леммы данного слова. Эти леммы, дополненные некоторой другой информацией, нейронная сеть учится классифицировать, предлагая уже на этапе практической работы ту или

иную лемму в качестве модификации заданного слова для нормализации всего выражения. В качестве общей модели в данной системе использован так называемый «Dependency Parser» [21] из библиотеки AllenNLP (<https://allennlp.org/allennlp>). Данный парсер предсказывает наличие дуги соответствующего типа между двумя словами в дереве зависимостей при синтаксическом анализе предложения. В системе Д. Анастасьева этот парсер адаптирован для предсказания связи между данной словоформой и одной из ее лемм. В качестве BERT-модели, порождающей векторные представления слов, в этой системе была использована модель RuBERT.

Сама система состоит из двух основных частей:

- (1) BERT-эмбеддера, который строит векторные представления слов входного предложения, и
- (2) LSTM-сети с классификатором, которые предсказывают необходимые модификации слов для нормализации всего выражения.

На вход LSTM-сети с классификатором подается комбинированное представление слов входного предложения, которое содержит

- (1) векторное представление самого слова, полученного от BERT-эмбеддера,
- (2) указатели на все возможные леммы данного слова из словаря,
- (3) информацию о положении данного слова внутри отмеченного выражения (т.е. одну из меток B, I, O из BIO-формата)

и некоторую другую информацию.

Данная система, примененная к нормализации выражений общего вида, показала результаты в 97–98% точности нормализации. Наши эксперименты по нормализации выражений, представляющих собой упоминания симптомов в выбранных предметных областях, дали результат 98,5%.

### 3. Отождествление выражений

После этапа нормализации выражений, представляющих собой упоминания симптомов, выполняется этап отождествления – отнесения найденного симптома к группе однотипных симптомов. Например, в медицинских анамнезах часто встречаются следующие семантически схожие понятия: *высыпания на коже, кожные высыпания, мелко-папулезные высыпания, эритематозные пятна, лихенизация, эритематозные бляшки.*

Отнесение подобного вида выражений к одной группе с выбором некоторого стандартного (эталонного) выражения в качестве представителя всей группы называют задачей нормализации понятий (concept normalization) или связыванием сущностей (entity linking). В области медицинской информатики уже делались попытки создания словарей



(тезаурусов) эталонных названий заболеваний, симптомов, лекарств и т.д., которые бы служили классификаторами медицинских терминов. Наиболее известной и разработанной системой является онтология UMLS (Unified Medical Language System) [9]. Однако русскоязычный эквивалент ее терминологической части – тезаурус Medical Subject Headings – имеет объем только 2% от английского варианта.

В UMLS основным понятием является так называемый «уникальный идентификатор понятия» (CUI – concept unique identifier). Такого рода идентификаторы выбраны большей частью для названий болезней, частей тела, лекарств и в меньшей степени для симптомов, в частности потому, что «симптомы» является трудно формализуемым понятием. Кроме того, что такие классификации неполны, они обладают и другими недостатками. В них часто отсутствуют выражения, которые мы называем «симптомами-характеристиками» (см. раздел 1.1.1). Например, в UMLS отсутствуют симптомы *тоны сердца ясные*, *тоны сердца ритмичные*, а имеется только понятие *тоны сердца*.

Введенная нами классификация на «симптомы-отклонения» и «симптомы-характеристики» вместе с реализованным механизмом отождествления схожих выражений позволяет в данном случае автоматически отнести выражения *тоны сердца ритмичные*, *сердечные тоны звучные*, *сердечные тоны ясные* к группе симптомов-характеристик с эталонным представителем *тоны сердца ясные*, а выражения *аритмия*, *учащенное разбиение*, *тахикардия*, *перебои в работе сердца* к группе симптомов-отклонений с эталонным представителем *сердцебиение*.

Таким образом, мы разработали собственную систему базовых (эталонных) терминов для упоминаний симптомов в выбранной нами предметной области (аллергические и пульмонологические заболевания), а также для наиболее важных симптомов коронавирусной инфекции. Также все упоминания симптомов из нашего корпуса предложений были вручную расклассифицированы по эталонным группам. Например, в группу *Зев(отклонения)* были отнесены выражения *катаральные явления*, *зев умеренно гиперемирован*, *зев рыхлый*, *зев: задняя стенка разрыхлена*, *зуд в зеве* и др., а в группу *зев спокоен* были отнесены выражения *зев спокоен*, *зев без катаральных явлений*, *миндалины не увеличены*.

Общее количество групп составило более 60. Так, например, в их состав вошли группы *Затруднение дыхания*, *Затруднение носового дыхания*, *Бронхообструктивный синдром*, *Дыхание в легких (отклонения)*, *Кашель*, *Кожные процессы*. В качестве базового метода отнесения выделенного выражения к некоторой группе (упоминаний) симптомов была использована возможность построения векторного представления этого выражения с помощью BERT-модели. Метод основывается на том, что если подать

на вход нейронной сети (BERT-модели) заданное выражение, которое заключено между специальными токенами **CLS** и **SEP**:

*CLS зев умеренно гиперемирован SEP,*

то кроме векторных представлений отдельных слов выражения нейросеть после последовательного чтения каждого из слов перейдет в заключительное состояние, числовое представление которого считается векторным представлением всего выражения и которое будет помещено в позицию токена **CLS**.

В практической реализации, как и на этапе нормализации выражений, в качестве BERT-модели нами была использована универсальная модель для русского языка RuBERT. Предварительные ее испытания показали, что она хорошо отождествляет семантически близкие термины, в том числе из медицинской области:

<i>затрудненное дыхание, одышка</i>	0.973
<i>экзема, сыпь</i>	0.968

Приведенные числа являются значениями (косинусной) меры близости соответствующих выражений.

Сам алгоритм отнесения заданного выражения к той или иной группе однотипных симптомов состоит из следующих основных шагов:

- (1) прямая проверка вхождения заданного выражения в ту или иную группу симптомов путем сопоставления текстовых представлений заданного выражения с каждым симптомом из группы; если такое вхождение найдено, то с данным выражением связывается название найденной группы;
- (2) если прямая проверка оказалась неудачной, то вычисляется векторное представление заданного выражения, после чего для каждой группы симптомов находится средневзвешенное значение близости данного выражения к каждому из симптомов в этой группе:

$$m = \frac{m_1 + m_2 + \dots + m_N}{N},$$

где  $m_i$  есть значение (косинусной) меры близости заданного выражения с  $i$ -ым симптомом из группы;  $N$  – количество симптомов в группе (векторные представления симптомов в группах вычисляются заранее); далее выбирается группа симптомов, для которой значение  $m$  оказалось максимальным, но не меньшим, чем 0,75;

- (3) если максимальное значение  $m$  оказалось меньше, чем 0,75, то исходное выражение считается неотнесенным к какой-либо группе и рассматривается как отдельная новая группа.

Проведенные эксперименты с упоминаниями симптомов из указанной

выше предметной области дали результаты в 92–93% по отнесению симптома к той или иной группе.

Следует, однако, отметить, что на модели RuBERT значения близости пар выражений *⟨одышка, тошнота⟩* и *⟨одышка, рвота⟩* оказались очень высокими: 0,96–0,97. Это объясняется тем, что модель RuBERT была построена на корпусе универсальных текстов, в которых данные пары слов чаще всего встречались в одинаковых контекстах. Следовательно, чтобы получить векторные представления медицинских терминов, которые более точно представляли бы их смысл, необходимо построение BERT-модели на достаточно большом корпусе специализированных медицинских текстов, в которых данные слова встречаются в более разнообразных контекстах.

#### 4. Обзор смежных работ

В данном разделе мы даем обзор наиболее важных работ, посвященных извлечению симптомов из неструктурированных (ЕЯ) медицинских текстов и их использованию в системах поддержки принятия диагностических решений. Для русского языка количество таких исследований невелико – все они перечислены ниже. Для них мы также проводим сравнение с нашей работой.

В работе [15] представлена система принятия диагностических решений на основе анамнезов пациентов, записанных в свободном неструктурированном виде. Система реализована в научном подразделении SberMedAI Сбербанка России и имеет название «Умный помощник врача «ТОР-3». По информации пресс-службы Сбербанка система внедрена во всех поликлиниках для взрослых г. Москвы. В качестве текста, поступавшего на вход нейросети, использовалась конкатенация содержимого разделов «симптомы» и «анамнез» из описания визита пациента к врачу. Общее количество использованных для обучения описаний визитов составило первоначально около 4 млн, а позднее было доведено до 14 млн. В качестве модели нейронной сети использовалась незначительно модифицированная стандартная BERT-модель. Модификация заключалась в добавлении в качестве последнего слоя нейронной сети полносвязного классификатора.

Классификатор выдает результат в виде одного из 265 кодов заболеваний, выбранных из Международной классификации болезней МКБ-10. Размер входных текстов был ограничен 128 словами (или, более точно, токенами).

Результат работы системы – 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 составила 47,5%, а по трем диагнозам – 68%.

Недостатком этого подхода является то, что такая система рассматривает текст «целиком», не выявляя в нем смысловые единицы, такие как заболевания, симптомы, названия диагностических и лечебных процедур, лекарства и т.д., а также отношения между ними. Поэтому данная система поставит «диагноз» по любому тексту, в том числе и по такому, в котором отсутствуют упоминания каких-либо симптомов, или вообще не являющемуся медицинским. Кроме того, низкая точность работы системы не позволяет рассматривать ее в качестве полезной компоненты систем поддержки принятия клинических решений.

В работе [16] подчеркивается необходимость извлечения из медицинских текстов отдельных сущностей, используемых в процессе постановки диагноза как врачом-практиком, так и в системах поддержки принятия клинических решений. В данной работе решаются задачи извлечения из текстов

- (1) упоминаний заболеваний,
- (2) названий лекарств.

Кроме того, для заболеваний предложены средства извлечения их атрибутов, таких как степень тяжести (*Severity*), этап течения заболевания (*Course*), часть тела (*Body location*) и некоторых других. Авторы работы отмечают отсутствие аннотированных корпусов клинических текстов на русском языке и предлагают собственный аннотированный корпус. Этот корпус кратко описан выше в разделе 1.2.1.

Напомним, что в данном корпусе размечаются не только сущности, но и отношения между ними, тогда как в нашей работе цель состояла в разработке системы извлечения только упоминаний симптомов, которые представляют собой непрерывные фрагменты текста, для чего и был подготовлен соответствующий корпус. Заявленная авторами работы [16] точность извлечения составила

- (1) для заболеваний – 95,1%,
- (2) для названий лекарств – 84,3%.

В работе [17] представлен подход к извлечению так называемых «смешанных упоминаний симптомов» (*mixture symptom mentions*) из текстов на китайском языке, посвященных традиционной китайской медицине. Под смешанным упоминанием симптомов понимаются выражения, в которых описывается в общем случае несколько симптомов с характеристиками их тяжести и указанием частей тела, к которым эти симптомы относятся. Задача извлечения для таких выражений состоит в выделении сущностей,

которые в них упоминаются, и отношений между ними. В качестве сущностей в данной работе рассматриваются симптом (Symptom), степень тяжести (Severity) и часть тела (Area of the body). Эти сущности могут находиться в отношениях `located_at` (Symptom, Area of the body) и `is_a_description_of` (Symptom, Severity).

Авторами работы предложена объединенная модель обучения (joint learning model), в которой нейронная сеть обучается одновременному извлечению как сущностей, так перечисленных выше отношений между ними. Обучающий корпус состоял из 2,2 тыс. текстов с 73 тыс. упоминаемых сущностей. В качестве модели нейронной сети была использована стандартная BERT-модель для китайского языка, дообученная на текстах по традиционной китайской медицине. Точность распознавания сущностей и отношений между ними составила 82,5%.

Если в нашей собственной системе мы извлекаем непрерывные фрагменты текста, являющиеся описанием одного или нескольких симптомов, то в работе [17] вначале извлекаются отдельные структурные части таких упоминаний, а затем из этих частей собираются описания отдельных симптомов. Наш механизм обработки перечисления симптомов (см. раздел 1.1.2) частично пересекается с подходом авторов данной работы.

В работе [18] заявлено создание алгоритма выявления подозрения на COVID-19 для использования в симптом-чекерах и системах поддержки принятия врачебных решений на основе информации, извлекаемой из текстов на русском языке. Исходными данными для алгоритма являются симптомы и их значения, извлекаемые из протоколов врачебных осмотров. Общее количество извлекаемых симптомов составило 14. Они были разделены на 2 группы:

- (1) *заложенность в груди, слабость, боль в мышцах, диарея, головная боль, кашель, конъюнктивит, сатурация, температура;*
- (2) *спутанность сознания, одышка, аносмия, поражения кожи, кровохарканье.*

Первая группа симптомов извлекалась, аналогично нашему подходу, с помощью методов распознавания именованных сущностей. Точность извлечения этих симптомов составила от 82,6% до 97,43% в зависимости от конкретного симптома. Вторая группа симптомов извлекалась с помощью специально сконструированных правил, в частности, включающих в себя поиск по синонимам этих симптомов – например, поиск симптома *аносмия* включал в себя поиск выражений *снижение обоняния, потеря обоняния*.

Отметим, что в нашей системе все такие формы упоминания симптомов извлекаются обученной нейросетью на первом этапе с последующим отождествлением. Для второй группы симптомов показатели точности

извлечения не приведены. Также в статье [18] не отмечено, как извлекались значения таких симптомов как *температура*, *частота дыхания*, *частота сердечных сокращений*, хотя эти значения используются в самом алгоритме предсказания наличия COVID-19. Обучающее множество состояло из 11 243 описаний первичных врачебных осмотров с установленными диагнозами ОРВИ и пневмония, т.е. данные документы не относились непосредственно к COVID-19. Размеченный корпус недоступен для исследовательских целей.

В работе [19] представлены

- (1) англоязычный «COVID-19 Annotated Clinical Text (CACT) Corpus»,
- (2) средства извлечения симптомов с их значениями и
- (3) система предсказания наличия COVID-19 на основе извлеченных симптомов.

Корпус текстов состоит из 1472 клинических описаний, содержащих около 30 тыс. упоминаний диагнозов, результатов тестов и симптомов. Данный корпус является первым аннотированным корпусом с информацией про COVID-19. В качестве симптомов в нем размечены не только симптомы как таковые (*cough*, *shortness of breath*), но и само слово *COVID*. Для слова *COVID* размечены его аргументы типа *Test Status* и *Assertion*, где, в свою очередь, аргумент типа *Assertion* может принимать значения *present*, *absent*, *possible*, *hypothetical*, *not patient*. Для самих симптомов размечены аргументы типа *Assertion*, *Change*, *Severity*, *Anatomy* и др. В силу сходства задачи извлечения симптомов со значениями с задачей извлечения событий с аргументами, что отмечено выше во Введении, в данной работе выделение симптомов реализовано как извлечение событий путем дообучения 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.

Аналогично нашей работе в данной системе реализована нормализация выражений, описывающих симптомы, но этот этап отнесен уже к алгоритму предсказания наличия COVID-19. Нормализация проводилась с использованием подготовленной вручную таблицы, отображающей возможные выражения для каждого из симптомов в их нормализованные формы. Например, выражения *coughing*, *coughs*, *distress coughing*, *distressed coughing* отображались в слово *cough*. Стандартизованная терминология, например, из онтологии UMLS, в данной системе не использовалась.

В работе [12] представлен русскоязычный аннотированный корпус RuCCoN, имеющий в своей основе корпус, описанный выше в разделе 1.2.1 с дополнительной разметкой, которая состоит в приписывании каждой из 16 тыс. сущностей, отмеченных в исходном корпусе, одного из 2,4 тыс.

уникальных терминов (концептов) из русскоязычной части онтологии UMLS. Новый корпус и руководство по аннотированию размещены в открытом доступе. В корпусе RuCCoN выделено 10 сущностей (семантических типов):


- (1) заболевание или синдром,
- (2) часть тела, орган,
- (3) признак или симптом,

и другие, но симптомы составляют только 6% от всех семантических типов. Отличие данной работы от нашей состоит в том, что авторы работы [12] для привязывания словоформы выделенной сущности к эталонному термину используют отдельную нейронную сеть, обученную такому связыванию (*entity linking*). Мы же использовали возможность получения векторного представления (эмбединга) произвольных выражений с помощью некоторой BERT-модели и определения их близости с помощью стандартной косинусной меры. Для разного вида тестовых множеств авторы работы [12] сообщают о точности связывания с помощью нейросети в пределах 52–58%.

## 5. Выводы и направления дальнейшей работы

В данной статье была представлена система извлечения упоминаний симптомов из медицинских текстов на русском языке. Состав извлекаемых симптомов был ограничен аллергическими и пульмонологическими заболеваниями, а также коронавирусной инфекцией. Для обучения нейросети извлечению симптомов был подготовлен аннотированный корпус объемом в 1,1 тыс. предложений. Помимо извлечения упоминаний симптомов в системе предусмотрены их нормализация и отождествление.

Точность извлечения составила 81%, нормализации – 98%, отождествления – 92%. Для реализации последних двух операций была использована модель RuBERT, но опыт ее использования показал, что для дальнейшего повышения качества работы нейросетей необходимы BERT-модели, построенные на основе больших специализированных корпусов медицинских русскоязычных текстов. Дальнейшая работа будет заключаться в переходе к решению задачи извлечения информации о симптомах в виде пар *⟨название симптома, значение симптома⟩*, а также в расширении предметной области на симптомы кардиологических заболеваний.

Авторы статьи выражают благодарность ЦКП «Центр данных ДВО РАН»  [22] за предоставление вычислительных ресурсов на базе GPU.

## Список литературы

- [1] Sutton R. T., Pincock D., Baumgar D. C., Sadowski D. C., Fedorov R. N., Kroeker K. I. *An overview of clinical decision support systems: benefits, risks, and strategies for success* // npj Digit. Med.– 2020.– Vol. **6**.– No. 3.– id. 17. doi ↑96
- [2] Kwan J. L., Lo L., Ferguson J., Goldberg H., Diaz-Martinez J. P., Tomlinson G., Grimshaw J. M., Shojania K. G. *Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials* // BMJ.– 2020.– Vol. **370**.– id. m3216. doi ↑96
- [3] Sha L., Qian F., Chang B., Sui Zh. *Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction*, Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) // Proceedings of the AAAI Conference on Artificial Intelligence.– 2018.– Vol. **32**.– No. 1.– Pp. 5916–5923. doi ↑97
- [4] Smirnova A., Cudre-Mauroux Ph. *Relation extraction using distant supervision: A survey* // ACM Computing Surveys.– 2019.– Vol. **51**.– No. 5.– id. 106.– 35 pp. doi ↑97, 104
- [5] Le Th. A., Burtsev M. S. *A deep neural network model for the task of named entity recognition* // International Journal of Machine Learning and Computing.– 2019.– Vol. **9**.– No. 1.– Pp. 8–13. URL doi ↑98, 107
- [6] Ji Z., Wei Q., Xu H. *BERT-based ranking for biomedical entity normalization*, AMIA Jt Summits Transl Sci Proc.– 2020.– Pp. 269–277. doi arXiv:1908.03548 ↑98
- [7] Anastasyev D. G. *Annotated span normalization as a sequence labelling task*, Papers from the Annual International Conference “Dialogue” (2021), Computational Linguistics and Intellectual Technologies.– vol. **20**.– 2021.– ISBN 978-5-7281-3032-1.– Pp. 8–15. doi ↑99, 109
- [8] Anastasyev D. G. *Exploring pretrained models for joint morpho-syntactic parsing of Russian*, Papers from the Annual International Conference “Dialogue” (2020), Computational Linguistics and Intellectual Technologies.– vol. **19**.– 2020.– ISBN 978-5-7281-3032-1.– Pp. 1-12. doi ↑109
- [9] Bodenreider O. *The Unified Medical Language System (UMLS): Integrating biomedical terminology* // Nucleic Acids Res.– 2004.– Vol. **32**, suppl. 1.– Pp. D267–D270. doi ↑99, 111
- [10] Coletti M. H., Bleich H. L. *Medical subject headings used to search the biomedical literature* // J. Am. Med. Inform. Assoc.– 2001.– Vol. **8**.– No. 4.– Pp. 317–323; Erratum in: J. Am. Med. Inform. Assoc.– 2001.– Vol. **8**.– No. 6.– Pp. 597. doi doi ↑99
- [11] Бледжянц Г. А., Исакова Ю. А., Осипов А. А. *Апробация и внедрение эффективного использования инструментов объединенной базы медицинских знаний системой дистанционного образования инновационных субъектов* // Человеческий капитал.– 2020.– № S12-1.– С. 199–205. \* ↑99
- [12] Nesterov A., Zubkova G., Miftahutdinov Z., Kokh V., Tutubalina E., Shelmanov A., Alekseev A., Avetisian M., Chertok A., Nikolenko S. *RuCCoN: Clinical concept normalization in Russian*, Findings of the Association for Computational Linguistics: ACL 2022 (Dublin, Ireland).– 2022.– Pp. 239–245. doi ↑116, 117



- [13] *Временные методические рекомендации Министерства здравоохранения Российской Федерации «Профилактика, диагностика и лечение новой коронавирусной инфекции (COVID-19)»*, Версия 14 (27.12.2021).– Министерство здравоохранения Российской Федерации.– 233 с. [URL](#) ↑<sup>105</sup>
- [14] *Краткое руководство по разметке тестового корпуса. Задача «Medicine light»*, Версия 1.6.– ИСА РАН и ИЦЗД.– 2014. [URL](#) ↑
- [15] Blinov P., Avetisian M., Kokh V., Umerenkov D., Tuzhilin A. *Predicting clinical diagnosis from patients electronic health records using BERT-based neural networks*, AIME 2020: Artificial Intelligence in Medicine, Lecture Notes in Computer Science.– vol. **12299**, eds. M. Michalowski, R. Moskovitch, Cham: Springer.– 2020.– ISBN 978-3-030-59136-6.– Pp. 111–121. [doi](#) ↑<sup>113</sup>
- [16] Shelmanov A. O., Smirnov I. V., Vishneva E. A. *Information extraction from clinical texts in Russian*, Papers from the Annual International Conference “Dialogue” (2015), Computational Linguistics and Intellectual Technologies.– vol. **14**.– 2015.– Pp. 560–572. [URL](#) ↑<sup>114</sup>
- [17] Sun Yu., Zhao Zh., Wang Zh., He H., Guo F., Luo Yu., Gao Q., Wei N., Liu J., Li G.-Zh., Li Z. *Leveraging a joint learning model to extract mixture symptom mentions from traditional Chinese medicine clinical notes* // BioMed Research International.– Vol. **2022**, Conference Issue: Big Data for Biomedical Research.– id. 2146236. [doi](#) ↑<sup>114, 115</sup>
- [18] 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. [doi](#) \* ↑<sup>105, 115, 116</sup>
- [19] 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. [doi](#) ↑<sup>105, 116</sup>
- [20] Zolotukhin D., Smurov I. *RuNormAS-2021: A shared task on Russian normalization of annotated spans*, Papers from the Annual International Conference “Dialogue” (2021), Computational Linguistics and Intellectual Technologies.– vol. **20**.– 2021.– ISBN 978-5-7281-3032-1.– Pp. 1245–1250. [doi](#) ↑<sup>109</sup>
- [21] Dozat T., Manning C. D. *Deep biaffine attention for neural dependency parsing*.– 2017.– 8 pp. [doi](#) [arXiv](#) 1611.01734 ↑<sup>110</sup>
- [22] Сорокин А. А., Макогонов С. В., Королев С. П. *Информационная инфраструктура для коллективной работы ученых Дальнего Востока России* // Научно-техническая информация. Сер. 1: Организация и методика информационной работы.– 2017.– № 12.– С. 14–16. \* ↑<sup>117</sup>


Поступила в редакцию 26.12.2022;  
одобрена после рецензирования 29.01.2023;  
принята к публикации 29.01.2023;  
опубликована онлайн 17.02.2023.

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

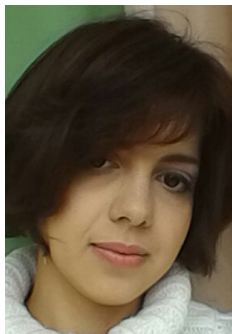
**e-mail:** [Yuri@serdyuk.botik.ru](mailto:Yuri@serdyuk.botik.ru)

**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

 0000-0002-7843-6870

**e-mail:** [nathalie.vlassova@gmail.com](mailto:nathalie.vlassova@gmail.com)

**Седа Рубеновна Момот**

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

 0000-0002-6097-6545

**e-mail:** [morlot@mail.ru](mailto:morlot@mail.ru)

*Все авторы сделали эквивалентный вклад в подготовку публикации.  
Авторы заявляют об отсутствии конфликта интересов.*



# A system for extracting symptom mentions from texts by means of neural networks

Yuri Petrovich **Serdyuk**<sup>1</sup>, Natalia Aleksandrovna **Vlasova**<sup>2</sup>,  
Seda Rubenovna **Momot**<sup>3</sup>

Ailamazyan Program Systems Institute of RAS, Ves'kovo, Russia

<sup>2</sup>[nathalie.vlassova@gmail.com](mailto:nathalie.vlassova@gmail.com)















**Abstract.** This paper presents a system for extracting symptom mentions from medical texts in natural (Russian) language. The system finds symptom mentions in texts, brings them to a standard form and identifies the found symptom to a group of similar symptoms. For each stage of processing we use a separate neural network. We extract symptoms of three areas of diseases: allergic and pulmonological diseases, as well as coronavirus infection (COVID-19). We present and describe an annotated corpus of sentences that is used to train neural networks for extracting symptom mentions. These sentences were marked up with the help of a simple XML-like language. An extended BIO-markup format was proposed for the sentences directly received at the input of the neural network. We give the quality evaluation of the symptom extraction accuracy under strict and flexible testing. Possible approaches to normalization and identification of symptom mentions and their implementation are described. Our results are compared with those achieved in similar researches, thus we show the place of our system among clinical decision support systems. (*In Russian*).

**Key words and phrases:** natural language processing, neural networks, information extraction, symptom mentions, annotated corpus, BERT-models, Covid-19

2020 *Mathematics Subject Classification:* 68T07; 68T50

**For citation:** Yuri P. Serdyuk, Natalia A. Vlasova, Seda R. Momot. A system for extracting symptom mentions from texts by means of neural networks. Program Systems: Theory and Applications, 2023, **14**:1(56), pp. 95–123. (*In Russ.*). [https://psta.psiras.ru/read/psta2023\\_1\\_95-123.pdf](https://psta.psiras.ru/read/psta2023_1_95-123.pdf)

## References

- [1] R. T. Sutton, D. Pincock, D. C. Baumgar, D. C. Sadowski, R. N. Fedorak, K. I. Kroeker. “An overview of clinical decision support systems: benefits, risks, and strategies for success”, *npj Digit. Med.*, **6**:3 (2020), id. 17. 
- [2] J. L. Kwan, L. Lo, J. Ferguson, H. Goldberg, J. P. Diaz-Martinez, G. Tomlinson, J. M. Grimshaw, K. G. Shojania. “Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials”, *BMJ*, **370** (2020), id. m3216. 
- [3] L. Sha, F. Qian, B. Chang, Zh. Sui. “Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction”, Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**:1 (2018), pp. 5916–5923. 
- [4] A. Smirnova, Ph. Cudre-Mauroux. “Relation extraction using distant supervision: A survey”, *ACM Computing Surveys*, **51**:5 (2019), id. 106, 35 pp. 
- [5] Th. A. Le, M. S. Burtsev. “A deep neural network model for the task of named entity recognition”, *International Journal of Machine Learning and Computing*, **9**:1 (2019), pp. 8–13.  
- [6] Z. Ji, Q. Wei, H. Xu. “BERT-based ranking for biomedical entity normalization”, *AMIA Jt Summits Transl Sci Proc.*, 2020, pp. 269–277.   1908.03548
- [7] D. G. Anastasyev. “Annotated span normalization as a sequence labelling task”, Papers from the Annual International Conference “Dialogue” (2021), *Computational Linguistics and Intellectual Technologies*, vol. **20**, 2021, ISBN 978-5-7281-3032-1, pp. 8–15. 
- [8] D. G. Anastasyev. “Exploring pretrained models for joint morpho-syntactic parsing of Russian”, Papers from the Annual International Conference “Dialogue” (2020), *Computational Linguistics and Intellectual Technologies*, vol. **19**, 2020, ISBN 978-5-7281-3032-1, pp. 1-12. 
- [9] O. Bodenreider. “The Unified Medical Language System (UMLS): Integrating biomedical terminology”, *Nucleic Acids Res.*, **32**, suppl. 1 (2004), pp. D267–D270. 
- [10] M. H. Coletti, H. L. Bleich. “Medical subject headings used to search the biomedical literature”, *J. Am. Med. Inform. Assoc.*, **8**:4 (2001), pp. 317–323; Erratum in: *J. Am. Med. Inform. Assoc.*, **8**:6 (2001), pp. 597.  
- [11] G. A. Bledzhyancz, Yu. A. Isakova, A. A. Osipov. “Approbation and implementation of the effective use of the tools of the integrated medical knowledge base by the system of distance education of innovative disciplines”, *Chelovecheskij kapital*, 2020, no. S12-1, pp. 199–205 (in Russian).
- [12] A. Nesterov, G. Zubkova, Z. Miftahutdinov, V. Kokh, E. Tutubalina, A. Shelmanov, A. Alekseev, M. Avetisian, A. Chertok, S. Nikolenko. “RuCCoN: Clinical concept normalization in Russian”, Findings of the Association for Computational Linguistics: ACL 2022 (Dublin, Ireland), 2022, pp. 239–245. 

- [13] *Vremennye metodicheskie rekomendacii Ministerstva zdravooxraneniya Rossijskoj Federacii «Profilaktika, diagnostika i lechenie novoj koronavirusnoj infekcii (COVID-19)»*, Versiya 14 (27.12.2021), Ministerstvo zdravooxraneniya Rossijskoj Federacii, 233 pp. [URL](#)
- [14] *Kratkoe rukovodstvo po razmetke testovogo korpusa. Zadacha «Medicine light»*, Versiya 1.6, ISA RAN i NCzZD, 2014. [URL](#)
- [15] P. Blinov, M. Avetisian, V. Kokh, D. Umerenkov, A. Tuzhilin. “Predicting clinical diagnosis from patients electronic health records using BERT-based neural networks”, AIME 2020: Artificial Intelligence in Medicine, Lecture Notes in Computer Science, vol. **12299**, eds. M. Michalowski, R. Moskovitch, Springer, Cham, 2020, ISBN 978-3-030-59136-6, pp. 111–121. [doi](#)
- [16] A. O. Shelmanov, I. V. Smirnov, E. A. Vishneva. “Information extraction from clinical texts in Russian”, Papers from the Annual International Conference “Dialogue” (2015), Computational Linguistics and Intellectual Technologies, vol. **14**, 2015, pp. 560–572. [URL](#)
- [17] Yu. Sun, Zh. Zhao, Zh. Wang, H. He, F. Guo, Yu. Luo, Q. Gao, N. Wei, J. Liu, G. -Zh. Li, Z. Li. “Leveraging a joint learning model to extract mixture symptom mentions from traditional Chinese medicine clinical notes”, *BioMed Research International*, **2022**, Conference Issue: Big Data for Biomedical Research, id. 2146236. [doi](#)
- [18] D. V. Gavrilov, A. V. Kirilkina, L. M. Serova. “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 informacionnye texnologii*, 2020, no. 4, pp. 51–58 (in Russian). [doi](#)
- [19] 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. [doi](#)
- [20] D. Zolotukhin, I. Smurov. “RuNormAS-2021: A shared task on Russian normalization of annotated spans”, Papers from the Annual International Conference “Dialogue” (2021), Computational Linguistics and Intellectual Technologies, vol. **20**, 2021, ISBN 978-5-7281-3032-1, pp. 1245–1250. [doi](#)
- [21] T. Dozat, C. D. Manning. *Deep biaffine attention for neural dependency parsing*, 2017, 8 pp. [doi](#) [arXiv:1611.01734](#)
- [22] A. A. Sorokin, S. V. Makogonov, S. P. Korolev. “Information infrastructure for the collective work of scientists from the Russian Far East”, *Nauchno-texnicheskaya informaciya. Ser. 1: Organizaciya i metodika informacionnoj raboty*, 2017, no. 12, pp. 14–16 (in Russian).