Системный анализ в медицине и биологии

Методы обработки естественного языка для извлечения факторов риска инсульта из медицинских текстов*

В.В. Донитова[I], Д.А. Киреев[I], Е.В. Титова[I], А.А. Акимова[I]

[I] Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия

Аннотация. Своевременное выявление факторов риска такого социально значимого заболевания, как инсульт, важно для организации профилактики этой патологии. Выбор наиболее эффективных современных методов обработки текстов для автоматического извлечения информации о наличии факторов риска у пациентов из электронных медицинских карт может повысить качество оказания превентивной медицинской помощи. Вопросы такого рода в области обработки естественного языка (Natural Language Processing, NLP) называются задачами извлечения именованных сущностей (Named Entity Recognition, NER). Для решения данной задачи были использованы методы извлечения информации (Information Extraction, IE) о заболеваниях и состоянии здоровья, основанные на вручную созданных правилах, машинном обучении (Machine Learning, ML) и глубоком обучении (Deep Learning, DL). На собранных и размеченных экспертами данных были проведены сравнительные экспериментальные исследования перечисленных методов. В экспериментах рассматривались 6 сущностей, однако описанные подходы и методы могут быть использованы для извлечения любых сущностей. По результатам экспериментов были сделаны выводы об эффективности разработанных методов и используемых текстовых характеристик для решения задачи.

Ключевые слова: *факторы риска, обработка естественного языка, извлечение именованных сущностей, машинное обучение, глубокое обучение.*

DOI: 10.14357/20790279210410

Введение

По данным Всемирной Организации Здравоохранения инсульт находится в тройке лидеров среди заболеваний, приводящих к смерти и инвалидности [1,2]. В силу того, что инсульты бывают разных типов, и их развитие при разном сочетании факторов риска протекает по-разному, как показали A.K. Boehme и соавт. [3], для проведения эффективной профилактики инсульта и снижения заболеваемости необходим анализ факторов риска.

Медицинские учреждения генерируют большой объем неструктурированных текстов, содержащих важную информацию о здоровье пациентов. К ним относятся анамнезы, результаты осмотров, описания результатов лабораторных и инструментальных исследований и др. В данной работе предлагается с использованием методов извлечения информации из клинических текстов вы-

* Работа выполнена при финансовой поддержке РФФИ, грант № 19-29-01090 мк.

явить наличие ряда факторов риска у пациентов, перенесших инсульт. В работе [4] было рассмотрено применение правил. В продолжение этого исследования были проведены эксперименты на основе трех подходов для извлечения информации из электронных медицинских карт (ЭМК). Первый из этих методов основан на правилах, второй – на машинном обучении, а именно на CRF (Conditional Random Fields, или метод условных случайных полей), третий – на машинном обучении, а именно на глубоком, с архитектурой модели BERT [5]. Для всех методов применялось обучение с учителем. Были проведены сравнительные экспериментальные исследования методов на размеченном корпусе клинических записей и сделаны выводы об их применении.

1. Подходы к извлечению информации из клинических текстов

В работе I. Neamatullah и соавт. [6] описывается алгоритм, написанный на языке программирования Perl и состоящий из словарей, регулярных выражений и разных правил для поиска конфиденциальной информации в медицинских текстах, например, имен или дат. При тестировании данного алгоритма авторами были получены следующие результаты: recall = 0,967, precision = 0,749.

В статье P. Sondhi и соавт. [7] рассматриваются две популярные модели машинного обучения: SVM (Support Vector Machine, или метод опорных векторов) и CRF (условные случайные поля) для извлечения следующей информации о медицинском случае: физическое обследование (включая симптомы) и курс лечения. Для обучения был собран набор данных HealthBoards. При обучении CRF авторами были получены следующие результаты: precision 0,62, recall 0,69 и F1 0,66, а для класса курса лечения: precision 0,51, recall 0,34 и F1 0,41, с общей точностью 64,82%. При обучении SVM были получены следующие результаты: для класса физических обследований – precision 0,75, recall 0,72 и F1 0,46, а для класса курса лечения – precision 0,54, recall 0,4 и F1 0,46, с общей точностью 71,69%.

Комбинированный подход описан в публикации H. Nayel и соавт. [8], в которой используются ансамбли классификаторов и разные модели представления сегментов для улучшения результатов работы методов извлечения информации из медицинских текстов. Лучший результат со значением метрики F1, равным 77,6, показала модель, состоящая из ансамблей.

В работе A. Arbabi и соавт. [9] используется модель Neural Concept Recognizer (нейронный распознаватель концепций), которая основана на сверточных сетях (Convolutional Neural Networks). Модель находит расстояние между эмбеддингом, представлением слова в виде числового вектора, текста или слова и эмбеддингом термина из разных онтологий. В данной работе рассматриваются две онтологии: Human Phenotype Ontology (HPO) и Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). Их иерархическая структура используется как неявное начальное значение для эмбеддингов терминов. Самые лучшие результаты в данной работе показала модель, основанная на онтологии HPO и обученная на аннотированных абстрактах статей PubMed [10]: Micro-Precision 80,3, Micro-Recall 62,4, Micro-F1-score 70,2 и Macro-Precision 80,5, Macro-Recall 68,2, Macro-F1-score 73,9.

В обзорной статье [11] рассматриваются разные подходы к извлечению информации из медицинских текстов с использованием глубокого обучения для извлечения именованных сущностей и отношения между ними (Relation Extraction, REX). Авторы рассмотрели статьи с 2017 до 2020 гг. и пришли к выводу, что использование методов глубокого обучения для решения ранее описанных задач приносит наилучшие результаты и является наиболее перспективным подходом. Однако авторы указывают на зависимость результатов глубокого обучения от количества данных при обучении. Для уменьшения количества необходимой информации можно воспользоваться методами активного обучения, описанными A. Shelmanov и соавт. [12]. В данной работе авторам удалось за небольшое число итераций активного обучения улучшить работу нескольких популярных на то время моделей глубокого обучения.

В настоящее время для решения задачи извлечения именованных сущностей используются методы глубокого обучения, такие как описанные в статье J. Lee и соавт. [13], где применяется популярная модель BERT, обученная на большом количестве медицинских текстов и позволяющая при использовании трансферного обучения (Transfer Learning) добиться передовых результатов.

Похожий подход в работе L. Gligic и соавт. [14], где использовалась модель, основанная на рекуррентных нейронных сетях (Recurrent Neural Networks, RNN), которые были обучены на большом количестве не аннотированных медицинских текстов и с использованием трансферного обучения. Эта модель позволила добиться высоких результатов для решения задач извлечения именованных сущностей и отношений между ними.

2. Материалы исследования

Экспериментальное исследование проводилось на основе обработки данных обезличенных ЭМК пациентов Федерального медико-биологического агентства (ФНКЦ ФМБА). ЭМК отбирались по наличию подходящего диагноза и по количеству госпитализаций (не менее двух для каждого пациента). Материал включал данные 341 ЭМК, из которых 239 были использованы в качестве обучающей выборки и 102 – в качестве контрольной. Из базы ФНКЦ ФМБА был выгружен файл с обезличенными данными ЭМК. Для извлечения информации применялась ручная разметка.

Для осуществления разметки был использован веб-инструмент «BRAT» [15], в котором открывались текстовые файлы историй болезни, и затем в них выделялись искомые сущности. Сущности в разметке могли быть выделены как независимо, так и с обозначением связи между разными сущностями. На рис. 1 приведен пример разметки сущности «головокружение».

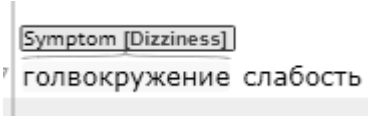

Рис. 1. Пример разметки отдельной сущности

На рис.2 приведен пример разметки сущности с одной связью. Основной сущностью «сахарный диабет» размечен диагноз, технической сущностью «значение сущности» размечен тип диабета, и обозначена связь («имеет значение») между основной и технической сущностью.

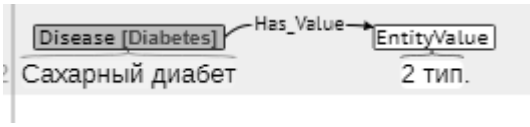

Рис. 2. Пример разметки сущности со значением

На рис. 3 приведен пример разметки сущности с несколькими связями. Основная сущность «гипертоническая болезнь» связана с тремя техническими сущностями «значение сущности».

В обрабатываемых текстах выделялись названия заболеваний с указанием таких характеристик, как стадия, степень, риск развития осложнений и др., симптомы, а также обозначались связи между используемыми сущностями и их характеристиками.

Кроме того, были созданы технические сущности для разметки характеристик и свойств основных сущностей, таких как локализация, время, стадия, степень, тип и т.д. К примеру, техническая сущность «значение сущности» (EntityValue) использовалась для записи стадии и степени гипертонической болезни, а также аналогичных характеристик других основных, т.е. соответствующих факторам риска инсульта, сущностей.

Сущности «гипертоническая болезнь» в эпикризах соответствовали следующие записи: артериальная гипертензия, артериальная гипертония, гипертоническая болезнь, гипертензивная болезнь, эссенциальная гипертензия, первичная гипертензия, АГ, ГБ, гипотензивная терапия, антигипертензивная терапия, принимает гипотензивные препараты, прием гипотензивных препаратов, регулярный прием гипотензивных препаратов, принимает антигипертензивные препараты. С сущностью «гипертоническая болезнь» также были связаны следующие характеристики: стадия, степень и риск сердечно-сосудистых осложнений (размечали как значения сущности).

Сущности «сахарный диабет» соответствовали следующие записи: сахарный диабет, СД, диагноз сахарного диабета, диабет. Значениями сущности размечали тип сахарного диабета.

Сущности «аритмия» соответствовали следующие записи: аритмия, аритмический синдром, аритмический вариант (ишемической болезни сердца), постоянная форма фибрилляции предсердий, фибрилляция предсердий, пароксизм аритмии, аритмическая кардиомиопатия, постинцизионное истмус-зависимое типичное трепетание предсердий, тахисистолия, персистирующая форма фибрилляции предсердий, фибрилляция и трепетание предсердий, тахистолическая форма типичного трепетания предсердий, пароксизмальная форма фибрилляции предсердий, пароксизмальная форма ФП, пароксизмы ФП, синдром нарушения ритма сердца, НРС, постоянная форма мерцательной аритмии, наджелудочковая экстрасистолия, желудочковая, суправентрикулярная экстрасистолия, предсердные экстрасистолы, ЖЭС, синусовая

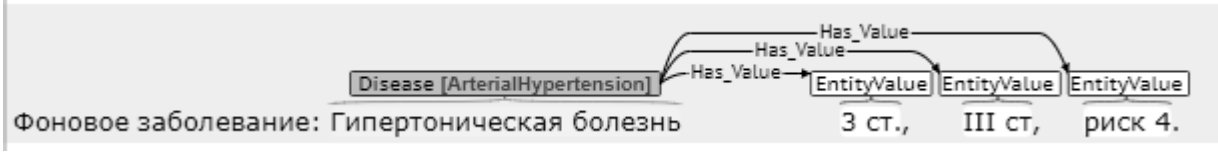

Рис. 3. Пример разметки сущности с несколькими значениями

аритмия, эпизод бигеминии, предсердно-желудочковая блокада, пароксизмальная СВТ, брадикардия, эпизоды наджелудочковой тахикардии. Значением сущности размечали степени блокад.

Сущности «дисциркуляторная энцефалопатия» соответствовали следующие записи: дисциркуляторная энцефалопатия, ДЭП, СМН 3 (III) ст., сосудисто-мозговая недостаточность 3ст., дисциркуляторная энцефалопатия.

Сущности «головокружение» соответствовали следующие записи: головокружение, вестибуло-атактический синдром, вестибулоатактический синдром, вестибуло-мозжечковый синдром, вестибуломозжечковый синдром, вестибулоатаксический синдром, вестибуло-атаксический синдром, вестибуло-кохлеарный синдром, вестибулокохлеарный синдром, вестибулопатия.

Сущности «стеноз сосудов» соответствовали следующие записи: стеноз, со стенозированием, стенозирование, стенозирован, стенозирующее поражение, окклюзионно-стенотическое поражение, окклюзия. Значением сущности размечали процент сужения просвета, а технической сущностью «локализация» размечали записи о том, какие именно сосуды поражены.

В процессе разметки была аннотирована 341 медицинская карта, в которых были выделены сущности, представленные в табл. 1. Было выделено 6 основных сущностей – гипертоническая болезнь, дисциркуляторная энцефалопатия, сахарный диабет, аритмия, головокружение и стеноз сосудов. В общей сложности коллекция размеченных терминов содержала 1368 включений по основным сущностям и 3618 включений по техническим.

Табл. 1

Размеченные сущности и количество их включений в текстах ЭМК

Название сущности	Количество включений
Гипертоническая болезнь	613
Сахарный диабет	199
Аритмия	153
Дисциркуляторная энцефалопатия	148
Стеноз сосудов	150
Головокружение	105

3. Методы извлечения информации

Извлечение информации из клинических текстов производилось с помощью трех методов: метода, основанного на правилах из регулярных выражений и газетира (Gazetteer); метода, основанного на условных случайных полях (CRF); метода на основе BERT.

3.1. Предобработка

Для корректной работы методов была выполнена предварительная обработка результатов разметки с целью их очистки от пробелов и служебных символов. Это позволило исключить влияние такого рода символов на конечный результат работы методов извлечения именованных сущностей.

Для возможности применения методов машинного обучения исходный набор данных был преобразован из формата разметки BRAT в модифицированную версию популярного для NLP формата CoNLL, в котором рассматриваются не символы, а слова, в то время как значения разделены специальным символом TAB, а предложения разделены пустой строкой.

3.2. Метод, основанный на правилах

Первый использованный в работе метод, основанный на правилах, был создан из коллекции размеченных терминов. Он состоит из газетира, инструмента для поиска сущностей в тексте, и регулярных выражений, которые и являются правилами.

Данный метод был реализован на языке программирования python. Реализация газетира была взята из библиотеки для обработки русского языка yargy [16]. Реализация регулярных выражений – из стандартной библиотеки python. Газетир применяется, чтобы отобрать записи, в которых используется термин. Это необходимо вследствие того, что поиск с помощью регулярных выражений очень ресурсоемок. Если газетир нашел используемый термин в документе, то по нему производится поиск с помощью регулярного выражения. Результат поиска выглядит как список спанов – непрерывных фрагментов текста, характеризуемых их положением, измеренным в символах, с извлеченным значением. Далее терминам из результата работы газетира присваиваются близко находящиеся значения из результата работы регулярных выражений, при этом расстояние между термином и значением задается как гиперпараметр.

В качестве экспериментальной сущности рассмотрим «гипертоническую болезнь». На основе размеченных данных можно увидеть, что данный термин может встречаться в следующих формулировках:

1) эссенциальная [первичная] гипертензия;
2) эссенциальная гипертензия;
3) артериальная гипертензия;
4) артериальная гипертония;
5) гипертензивная [гипертоническая] болезнь;
6) гипертензивная болезнь;
7) гипертоническая болезнь;

8) гипертонической болезни сердца

Также для повышения точности были рассмотрены популярные опечатки:

1) гипретоническая болезнь;
2) гипертончиеская болезнь;
3) гиперточниеская болезнь;
4) гипретоническая болезнь;
5) гипертончиеская болезнь.

Газетир нормализует данные формулировки и производит поиск по тексту. Если газетир находит термин, то по тексту производится поиск с помощью регулярных выражений, которые извлекают характеристику термина, такую как степень, стадия или риск. Далее с использованием гиперпараметра термин и характеристика связываются в тех случаях, если минимальное расстояние между спаном термина и характеристики меньше заданного гиперпараметра.

3.3. Метод условных случайных полей

Второй используемый в работе метод основан на машинном обучении, а именно на методе условных случайных полей (CRF). Модель условных случайных полей является подклассом марковских случайных полей (MRF) и представляет графическую ненаправленную вероятностную дискриминантную модель. Данный метод отличается простотой архитектуры при низкой скорости обучения, однако он менее эффективен при работе с более сложными сущностями. Формальное определение моделей CRF приведено в [17].

Для реализации данного метода использовался язык программирования python и популярная библиотека для машинного обучения sklearn, которая была представлена в работе [18]. Для реализации метода, основанного на CRF, был применен аддон для sklearn: sklearn-crfsuite [19] основанный на python-crfsuit'e [20], который является привязкой (binding) к библиотеке CRFsuite [21]. При обучении были подобраны гиперпараметры для L1 и L2 регуляризации.

3.4. Применение глубокого обучения

Третий использованный в работе метод основан на глубоком обучении, а именно на BERT [5]. В данной реализации использовался предварительно обученный на русском языке BERT – RuBERT [22]. Изначально BERT был обучен на задачах «Masked Language Model» и «Next Sentence Prediction», поэтому, чтобы использовать BERT для решения данной задачи, необходимо было дообучить имеющуюся модель на задаче извлечения именованных сущностей без изменения архитектуры. Для этого использовался язык программирования python и библиотека HuggingFace [23]. Визуализацию данной архитектуры можно увидеть на рис. 4.

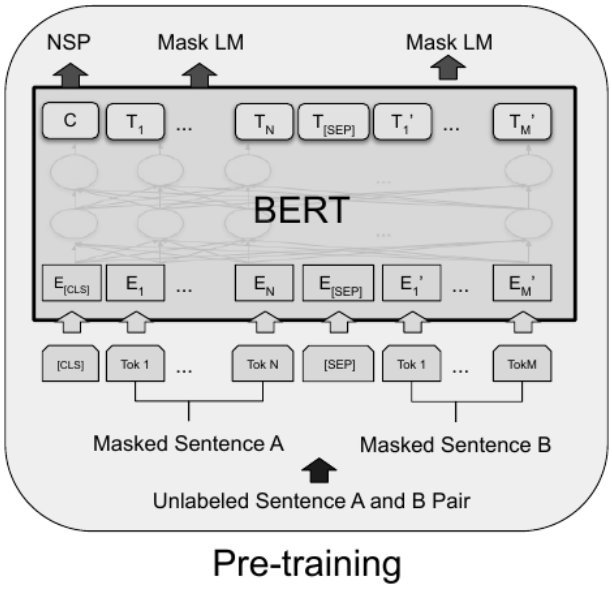

Рис. 4. Визуализация архитектуры модели BERT [4]

4. Экспериментальное исследование методов

Для всех методов использовалось обучение с учителем, для чего набор данных был разделен на 70% данных, которые были использованы для обучения и 30% данных, которые были использованы для тестирования. Результаты были оценены по трем параметрам:

1) Precision – доля релевантных значений среди извлеченных значений. Данная метрика высчитывается по уравнению (1), где TP – количество правильно предсказанных тэгов, FP – количество неправильно предсказанных тэгов:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

2) Recall – доля извлеченных релевантных значений. Данная метрика высчитывается по уравнению (2), где TP – количество правильно предсказанных тэгов, FN – количество не предсказанных тэгов:

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (2)$$

3) F_1 – точность теста. Данная метрика высчитывается по уравнению (3), где TP – количество правильно предсказанных тэгов, FP – количество неправильно предсказанных тэгов, FN – количество не предсказанных тэгов:

$$\text{F1} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}. \quad (3)$$

В табл. 2 приведены оценки метода, основанного на правилах.

Табл. 2

Результаты оценки метода, основанного на правилах

Сущность	**Precision, %**	**Recall, %**	**F_1, %**
АН (гипертоническая болезнь)	96,5	82,2	88,8
Angiostenosis (стеноз сосудов)	83,0	96,5	89,3
Arrhythmia (аритмия)	88,7	75,5	81,5
DEP (дисциркуляторная энцефалопатия)	82,1	90,2	86,0
Diabetes (диабет)	85,5	96,0	91,5
Dizziness (головокружение)	83,7	90,4	87,0

В табл. 3 приведены оценки двух методов: метода, основанного на CRF и метода, основанного на глубоком обучении. Обучение данных методов осуществлялось на одном наборе данных.

При оценке метода, основанного на правилах, с помощью меры F1, были получены результаты от 81,5 до 91,5%. Данная оценка является ожидаемой для метода с вручную созданными правилами [6]. Метод, основанный на CRF, показал результаты от 23,0 до 83,5%, однако стоит заметить, что оценка F1 пропорциональна количеству сущностей в наборе данных и количеству форм записи сущности, поэтому при необходимости повышения результативности метода необходимо увеличить количество сущностей в наборе данных. Метод, основанный на глубоком обучении, показал результаты от 24,3 до 99,4%. Необходимо отметить более низкие, по сравнению с другими методами, результаты работы метода глубокого обучения, основанного на BERT (RuBERT), при извлечении сущности «гипертоническая болезнь».

Высокий recall свидетельствует о правильном выделении большинства форм заболевания. Однако пониженный precision может говорить о том, что модель избыточно выделяет похожие формы записи в ЭМК, например, «артериальное давление», которое при обработке текста воспринимается как «артериальная гипертония».

Заключение

Инсульт является второй по значимости причиной смерти и первой по значимости причиной тяжелой инвалидности во всем мире. Поэтому предотвращение инсульта является одной из важнейших целей современного здравоохранения. Многие давно известные факторы риска инсульта учитываются в клинической практике и часто успешно купируются, но их все увеличивающаяся распространенность и омоложение являются поводом для новой оценки с помощью анализа на большом массиве клинических данных.

Методы извлечения информации из текста можно применять для обработки ЭМК с целью выявления и анализа факторов риска.

В процессе исследования был подготовлен большой объем аннотированных ЭМК, которые использовались для создания правил, обучения моделей и их тестирования.

Экспериментально была продемонстрирована целесообразность использования для извлечения информации метода, основанного на правилах, и методов машинного обучения. Метод, основанный на правилах, показал преимущества при извлечении простых или малочисленных

Табл. 3

Результаты оценки метода, основанного на CRF и метода, основанного на глубоком обучении

Метод	**CRF**			**BERT**		
Сущность\Метрика	Precision, %	Recall, %	F_1, %	Precision, %	Recall, %	F_1, %
АН (гипертоническая болезнь)	83,5	84,0	83,5	70,4	91,2	79,5
Angiostenosis (стеноз сосудов)	33,0	27,5	29,5	99,8	99,0	99,4
Arrhythmia (аритмия)	56,0	28,5	37,5	90,3	97,2	93,6
DEP (дисциркуляторная энцефалопатия)	90,5	69,5	79,0	38,8	60,0	45,7
Diabetes (диабет)	63,0	71,5	66,5	30,2	77,7	43,5
Dizziness (головокружение)	28,5	19,0	23,0	15,3	60,0	24,3

сущностей, в то время как методы, основанные на машинном обучении, – для всех остальных случаев.

Разработанные методы извлечения информации из клинических текстов позволят ускорить и усовершенствовать процесс выявления факторов риска такого тяжелого заболевания, как острое нарушение мозгового кровообращения.

Создание системы автоматического анализа ЭМК на предмет выявления факторов риска хронических нарушений мозгового кровообращения позволит сформировать более эффективную систему профилактики инсультов.

Литература

1. *Johnson W., Onuma O., Owolabi M. and Sachdev S.* Sep. Stroke: a global response is needed. Bull. World Health Organ. 2016. Vol. 94. No. 9. P. 634-634A.
2. *Thrift A.G. et al.* Jan. Global stroke statistics. Int. J. stroke Off. J. Int. Stroke Soc. 2014. Vol. 9. No. 1. P. 6–18.
3. *Boehme A.K., Esenwa C. and M.S. V Elkind.* Feb. Stroke Risk Factors, Genetics, and Prevention. Circ. Res. 2017. Vol. 120. No. 3. P. 472–495.
4. *Благосклонов Н.А. и др.* Лингвистический анализ историй болезни для выявления факторов риска инсульта / Труды ИСА РАН. 2020. Т. 70. № 3. С. 75-85.
5. *Devlin J., Chang M.-W., Lee K. and Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.
6. *Neamatullah I. et al.* Automated de-identification of free-text medical records. BMC Med. Inform. Decis. Mak. 2008. Vol. 8. No. 1. P. 32.
7. *Sondhi P., Gupta M., Zhai C. and Hockenmaier J.* Shallow Information Extraction from Medical Forum Data. Coling 2010: Posters. 2010. P. 1158–1166. Available at: https://www.aclweb.org/anthology/C10-2133.
8. *Nayel H. and Shashirekha H.L.* "Improving {NER} for Clinical Texts by Ensemble Approach using Segment Representations," in Proceedings of the 14th International Conference on Natural Language Processing ({ICON}-2017). 2017. P. 197–204. Available at: https://www.aclweb.org/anthology/W17-7525.
9. *Arbabi A., Adams D.R., Fidler S. and Brudno M.* May Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. JMIR Med. informatics. 2019. Vol. 7. No. 2. P. e12596.
10. PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/ (дата обращения 15.04.2021).
11. *Hahn U. and Oleynik M.* Aug. Medical Information Extraction in the Age of Deep Learning. Yearb. Med. Inform. 2020. Vol. 29. No. 1. P. 208–220.
12. *Shelmanov A. et al.* Active Learning with Deep Pre-trained Models for Sequence Tagging of Clinical and Biomedical Texts. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019. P. 482–489.
13. *Lee J. et al.* Feb. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020. Vol. 36. No. 4. P. 1234–1240.
14. *Gligic L., Kormilitzin A., Goldberg P. and Nevado-Holgado A. Jan.* Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks. Neural Netw. 2020. Vol. 121. P. 132–139.
15. *Stenetorp P. et al.* BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012. P. 102-107.
16. Yargy: Rule-based facts extraction for Russian language. Available at: https://github.com/natasha/yargy (дата обращения 15.04.2021).
17. *Lafferty J.D., McCallum A. and Pereira F.C.N.* Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning. 2001. P. 282–289.
18. *Pedregosa F. et al.* Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011. Vol. 12. No. null. P. 2825–2830.
19. Sklearn-crfsuite: scikit-learn inspired API for CRFsuite. Available at: https://github.com/TeamHG-Memex/sklearn-crfsuite (дата обращения 15.04.2021).
20. Python-crfsuite: a python binding for crfsuite, Available at: https://github.com/scrapinghub/python-crfsuite (дата обращения 15.04.2021).
21. *Okazaki N.* 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). Available at: http//www. chokkan. org/software/crfsuite.
22. *Kuratov Y. and Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019.
23. *Wolf T. et al.* HuggingFace's Transformers: State-of-the-art Natural Language Processing. 2020.

Донитова Виктория Владимировна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Научный сотрудник. Количество печатных работ: 20. Область научных интересов: извлечение знаний, интеллектуальные системы, системы поддержки принятия решений, экспертные системы. E-mail: vdonitova@gmail.com (Ответственный за переписку).

Киреев Данил Алексеевич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Техник 2-ой категории. Количество печатных работ: 4. Область научных интересов: машинное обучение, глубокое обучение, активное обучение, обработка естественного языка, компьютерное зрение, извлечение именованных сущностей, программирование микроконтроллеров. E-mail: kireev@isa.ru

Титова Елизавета Викторовна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Инженер-исследователь. Количество печатных работ: 2. Область научных интересов: обработка естественного языка, проблемно-ориентированные системы, экспертные системы, медицинские информационные системы. E-mail: elz.titova@gmail.com

Акимова Анна Анатольевна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» г. Москва, Россия. Техник. Количество печатных работ: 1. Область научных интересов: экологические инновации, машинное обучение, обработка естественного языка, извлечение именованных сущностей. E-mail: anna.djerg@mail.ru

Natural language processing models for extraction of stroke risk factors from electronic health records

V.V. Donitova[I], D.A. Kireev[I], E.V. Titova[I], A.A. Akimova[I]

[I] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. High social impact of stroke makes early detection of stroke risk factors crucial for its prevention. It is important to use the most efficient natural language processing (NLP) methods for automatic extraction of information about risk factors from the electronic health records (EHRs) to improve the quality of preventive medical care.

The authors have developed methods to extract information about diseases and health status of patients based on manually created rules, statistical machine learning and deep learning to solve the problem of named entity recognition (NER) in clinical records. Comparative experimental studies of the developed methods were conducted on a marked-up corpus of clinical records. As a result, conclusions are made on the effectiveness of the developed methods.

Keywords: *risk factors, natural language processing, named entity recognition, machine learning, deep learning.*
DOI: 10.14357/20790279210410

References

1. *Johnson W., Onuma O., Owolabi M. and Sachdev S.* Sep. 2016. Stroke: a global response is needed. Bull. World Health Organ., vol. 94, no. 9, pp. 634-634A.
2. *Thrift A.G. et al.* Jan. 2014. Global stroke statistics. Int. J. stroke Off. J. Int. Stroke Soc., vol. 9, no. 1, pp. 6–18.
3. *Boehme A.K., Esenwa C. and M.S. V Elkind.* Feb. 2017. Stroke Risk Factors, Genetics, and Prevention. Circ. Res., vol. 120, no. 3, pp. 472–495.
4. *Blagosklonov N.A. et al.* 2020. Linguistic analysis of disease history for identifying stroke risk factors. Trudy Instituta sistemnogo analiza rossiyskoy akademii nauk" ("Proceedings of the Institute for Systems Analysis of the Russian Academy of Science"), vol. 70, no. 3, pp. 75-85.
5. *Devlin J., Chang M.-W., Lee K. and Toutanova K.* 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
6. *Neamatullah I. et al.* 2008. Automated de-identification of free-text medical records. BMC Med. Inform. Decis. Mak., vol. 8, no. 1, p. 32.
7. *Sondhi P., Gupta M., Zhai C. and Hockenmaier J.* 2010. Shallow Information Extraction from Medi-

cal Forum Data. Coling 2010: Posters, pp. 1158–1166. Available at: https://www.aclweb.org/anthology/C10-2133.
8. *Nayel H. and Shashirekha H.L.* "Improving {NER} for Clinical Texts by Ensemble Approach using Segment Representations," in Proceedings of the 14th International Conference on Natural Language Processing ({ICON}-2017), 2017, pp. 197–204. Available at: https://www.aclweb.org/anthology/W17-7525.
9. *Arbabi A., Adams D.R., Fidler S. and Brudno M.* May 2019. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. JMIR Med. informatics, vol. 7, no. 2, p. e12596,
10. PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/ (дата обращения 15.04.2021).
11. *Hahn U. and Oleynik M.* Aug. 2020. Medical Information Extraction in the Age of Deep Learning. Yearb. Med. Inform., vol. 29, no. 1, pp. 208–220.
12. *Shelmanov A. et al.* 2019. Active Learning with Deep Pre-trained Models for Sequence Tagging of Clinical and Biomedical Texts. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 482–489.
13. *Lee J. et al.* Feb. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, vol. 36, no. 4, pp. 1234–1240.
14. *Gligic L., Kormilitzin A., Goldberg P. and Nevado-Holgado A.* Jan. 2020. Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks. Neural Netw., vol. 121, pp. 132–139.
15. *Stenetorp P. et al.* 2012. BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Pp. 102-107.
16. Yargy: Rule-based facts extraction for Russian language. Available at: https://github.com/natasha/yargy (дата обращения 15.04.2021).
17. *Lafferty J.D., McCallum A. and Pereira F.C.N.* 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289.
18. *Pedregosa F. et al.* 2011. Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res., vol. 12, no. null, pp. 2825–2830.
19. Sklearn-crfsuite: scikit-learn inspired API for CRFsuite. Available at: https://github.com/Team-HG-Memex/sklearn-crfsuite (дата обращения 15.04.2021).
20. Python-crfsuite: a python binding for crfsuite, Available at: https://github.com/scrapinghub/python-crfsuite (дата обращения 15.04.2021).
21. *Okazaki N.* 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). Available at: http//www. chokkan. org/software/crfsuite.
22. *Kuratov Y. and Arkhipov M.* 2019.Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language.
23. *Wolf T. et al.* 2020.HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Donitova V.V. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. E-mail: vdonitova@gmail.com

Kireev D.A. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. E-mail: kireev@isa.ru

Titova E.V. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. E-mail: elz.titova@gmail.com

Akimova A.A. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. E-mail: anna.djerg@mail.ru