https://doi.org/10.26160/2474-5901-2023-39-86-101

МОДЕЛИ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РАСПОЗНАВАНИЯ МЕДИЦИНСКИХ ИМЕНОВАННЫХ СУЩНОСТЕЙ

Вульфин А.М., Гаянова М.М., Улямаев Т.И.
Уфимский университет науки и технологий, Уфа, Россия

Ключевые слова: именованные сущности, глубокие нейронные сети, наборы данных, модели трансформеры, интеллектуальный анализ текстовых данных.
Аннотация. В работе проанализированы подходы к построению моделей распознавания медицинских именованных сущностей для повышения эффективности анализа слабоструктурированных текстовых данных, что является актуальной задачей для совершенствования систем поддержки принятия решений в клинической практике. Выполнен сравнительный обзор и классификация моделей глубоких нейронных сетей в задачах распознавания медицинских именованных сущностей. Проведенный эксперимент на корпусе медицинских текстов заданной тематики показал, что модели общего назначения не позволяют приблизиться по качеству выделения именованных сущностей к экспертной разметке. Применение технологий переноса обучения позволяет существенно повысить эффективность моделей трансформеров.

DEEP NEURAL NETWORK MODELS FOR MEDICAL NAMED ENTITY RECOGNITION

Vulfin A.M., Gayanova M.M., Ulyamaev T.I.
Ufa University of Science and Technology, Ufa, Russia

Keywords: named entities, deep neural networks, datasets, transformers models, text data mining.
Abstract. The paper analyzes approaches to constructing models for recognizing medical named entities to improve the efficiency of analyzing semi-structured text data, which is an urgent task for improving decision support systems in clinical practice. A comparative review and classification of deep neural network models in medical named entity recognition tasks has been performed. An experiment conducted on a corpus of medical texts on a given topic showed that general-purpose models do not allow the quality of identifying named entities to approach expert markup. The use of transfer learning technologies can significantly increase the efficiency of transformer models.

Введение

Постоянный рост объема неструктурированной и/или слабоструктурированных данных в виде текста на естественном языке и растущая потребность в извлечении из имеющихся данных структурированной информации являются актуальной проблемой [1]. В медицинских информационных системах накоплены значительные объемы слабоструктурированных текстовых данных, содержащих подробную клиническую информацию [2]. Например, данные из медицинских выписок могут быть полезны при анализе причинно-следственных связей, статистическом анализе и поиске скрытых закономерностей течения заболевания. Непосредственное использование исходных медицинских текстов без предобработки и структурирования информации невозможно [3].

Например, методы извлечения информации из текста на естественном языке можно применять для обработки электронных медицинских карт (ЭМК) с целью выявления и анализа факторов риска развития или течения заболеваний. В работе [3] показано, что создание системы автоматического анализа ЭМК для выявления факторов риска хронических нарушений мозгового кровообращения позволит сформировать более эффективную систему профилактики инсультов.

Распознавание именованных сущностей (Named Entity Recognition, NER) является одной из основных технологий обработки текстовой информации на естественном языке (Natural Language Processing, NLP) и важной частью поиска ценной информации в медицинских текстах. Технология NER позволяет выделять необходимую информацию в медицинских текстах и оказывать медицинскому персоналу значимую поддержку в принятии клинических решений, доказательной медицине и мониторинге эпидемических заболеваний [4]. NER может применяться во многих областях, включая поиск информации, построение систем ответов на вопросы, классификации текстов, извлечения связей и т.п. Извлеченные именованные сущности помогают понять тематическое содержание текста и определить ключевые слова текста.

1. Распознавание именованных сущностей в медицинских текстах

Системы NER в здравоохранении [5] позволяют обнаруживать и классифицировать такие объекты, как: имена пациентов, медицинские термины и различные термины, из неструктурированного текста. Например, клиническая система NER [6] идентифицирует и извлекает медицинские объекты, их контекст и взаимоотношения из больших объемов неструктурированных клинических данных с использованием нейросетевых NLP моделей глубокого обучения.

Категории, которые обычно обнаруживаются с помощью систем NER для медицинских текстов [6]:

– СОСТОЯНИЕ ЗДОРОВЬЯ: выявляет заболевания, травмы, симптомы или любые жалобы на здоровье.

– МЕДИКАМЕНТ: названия лекарств, методов лечения или других терапевтических веществ.

– АНАТОМИЯ: термины, относящиеся к частям тела, органам или анатомическим структурам.

– ПРОЦЕДУРА: обозначает медицинские вмешательства, тесты или операции.

– РЕЗУЛЬТАТ ИСПЫТАНИЙ: освещает результаты медицинских анализов.

– ЧЕЛОВЕК: определяет лиц, участвующих в уходе за пациентом или в личной жизни.

– TIME: идентифицирует ссылки, связанные со временем, такие как продолжительность, частота или конкретные даты.

Также распознавание именованных сущностей применяется для идентификации и последующего удаления/кодирования информации PII

(частная и конфиденциальная информацию, Personally Identifiable Information или PII) из отчетов независимой медицинской экспертизы (IME), подготовленных врачом [7]. NER представляет собой высокопроизводительную альтернативу ручной деидентификации, которая может быть реализована с помощью существующих инструментальных средств при сравнительно небольших затратах.

Распознавание именованных сущностей для языков с богатой морфологией [8] является сложной задачей, обусловленной богатством словоформ и неоднозначностью представления. Клинические медицинские тексты обладают существенной сложностью для автоматизированного анализа, например, в виду большого количества сокращений, опечаток и синонимов [9]. В статье [3] рассматриваются основные проблемы при извлечении данных из медицинских выписок пациентов с сердечно-сосудистыми заболеваниями. Выделяется три типа данных для извлечения: персональные данные, диагнозы, количественные характеристики диагнозов. В статье [10] представлена система для извлечения упоминаний симптомов из медицинских текстов на русском языке. Система осуществляет нахождение симптомов в тексте, их нормализацию (приведение к стандартной форме) и отождествление – отнесение найденного симптома к группе однотипных симптомов.

Таким образом, исследование и разработка моделей и алгоритмов NER для повышения эффективности анализа слабоструктурированных текстовых данных является актуальной задачей для совершенствования систем поддержки принятия решений в клинической практике.

2. Анализ существующих подходов к построению систем NER

Для создания систем NER в разное время были предложены различные подходы, начиная от достаточно простых, основанных на лексических правилах, и заканчивая более сложными нейросетевыми моделями (табл. 1).

В работе [14] представлен подход на основе построения словарей, регулярных выражений и экспертных правил для поиска конфиденциальной информации в медицинских текстах.

В работе [15] исследование проводилось на основе обработки данных обезличенных ЭМК пациентов Федерального медико-биологического агентства. Набор данных включал 341 ЭМК с обезличенными данными. Для извлечения информации применялась ручная разметка. Метод, основанный на правилах, показал преимущества при извлечении простых или малочисленных сущностей, в то время как методы, основанные на машинном обучении, – продемонстрировали преимущества для всех остальных случаев.

В статье [16] предложены две модели машинного обучения: SVM (Support Vector Machine, или метод опорных векторов) и CRF (условные случайные поля) для извлечения информации о медицинском обследовании (включая симптомы) и курсе лечения.

Табл. 1. Обзор существующих подходов к построению систем NER [11]

Подход	Область применения	Достоинства	Недостатки	Примеры
Группа: Классические системы				
Экспертные правила	извлечение классических именованных сущностей: имена (персоны), организации, местоположения с небольшой вариабельностью	простота конструирования и верификации правил;	сформировать исчерпывающий набор правил очень сложно и трудоемко; применение для ограниченной и строго определенной предметной области	rule-based системы; регулярные выражения; шаблоны капитализации [12]
Построение и использование онтологий	использование для узкоспециализированной предметной области	высокая точность распознавания именованных сущностей	отсутствие обобщающей способности и невозможность переноса на смежные предметные области	
Модели машинного обучения для классификации на основе выделенных признаков	Извлечение классических именованных сущностей	Автоматическое построение базы правил	сложность конструирования вектора признаков; высокая вероятность переобучения	деревья принятия решений; скрытые марковские цепи (HMM); машины опорных векторво; условные случайные поля (CRF) – вероятностные модели для структурированного предсказания
Группа: Нейросетевые и гибридные системы				
Модели на основе глубоких нейронных сетей	Извлечение широкого класса именованных сущностей с высокой степенью гетерогенности и вариабельности	Возможность построения гетерогенного вектора признаков, включая: морфологические, синтаксические и семантические признаки; Возможность извлечения репрезентативных признаков; Возможность получения контекстно-чувствительного представления из последовательности слов	Необходимость обширной и представительной обучающей выборки	нескольких двунаправленных рекурентных слоёв + классификатор; Long Short-Term Memory (LSTM); Двунаправленные рекуррентные нейронные сети (bi-LSTM); объединённая модель BI-LSTM и CRF; CharCNN-BLSTM-CRF [13]

Табл. 1. Продолжение

Подход	Область применения	Достоинства	Недостатки	Примеры
Группа: Системы на основе языковых моделей				
Семантические модели (в том числе, дистрибутивные) и глубокие контекстно-зависимые представления слов, взятые из обученной двунаправленной языковой модели, и их комбинация с дистрибутивными представлениями		Использование векторного представления на уровне отдельных символов, слов и фрагментов теста для семантической оценки близости описаний; глубокое понимание грамматики и других элементов естественного языка Предобученные на неразмеченных данных векторные представления слов; тонкая настройка двунаправленной языковой модели; глубокие контекстно-зависимые представления слов	Не ясна связь между обучением языковой модели и качеством выделения именованных сущностей	Двунаправленная языковая модель BiRNN [12]; Context2vec; Модели трансформеры

В статье [2] рассматривается построение дискриминационных моделей для решения задачи извлечения именованных сущностей из медицинских текстов на русском языке. Рассматриваются такие модели, как CRF и SVM. Применение этих методов к текстам на русском языке, а тем более к медицинским текстам, насыщенным специфической медицинской терминологией, по-прежнему остается проблемой. Для решения этой проблемы описываются процессы выделения признаков и построения моделей в контексте указанных текстов.

В работе [17] построены ансамбли классификаторов и модели представления сегментов для повышения эффективности извлечения информации из медицинских текстов.

В работе [18] предложена модель Neural Concept Recognizer (нейросетевой распознаватель концепций), которая основана на сверточных нейронных сетях (Convolutional Neural Networks). Модель находит расстояние между векторным вложением для отдельного слова или фрагмента текста и векторным вложением термина из предварительно построенных онтологий.

В работе [19] представлен метод использования статистических синтаксических анализаторов NER на медицинском корпусе русского языка. Предложенная модель показала хорошие результаты при извлечении

признаков из неструктурированных историй болезней и электронных медицинских карт.

В статье [20] проанализированы различные подходы к извлечению информации из медицинских текстов с использованием нейросетевых моделей глубокого обучения для извлечения именованных сущностей и отношений между ними (Relation Extraction, REX). Показано, что использование методов глубокого обучения позволяет добиться наилучших результатов в задачах NER с 2017 года. Отмечено, что качество работы подобных решений существенно зависит от объема обучающей выборки.

В статье [21] представлены результаты по классификации русскоязычных именованных сущностей и поиску эквивалентных именованных сущностей с использованием векторных вложений на уровне слов на уровне слов и словосочетаний. Показано, что вектор контекста слова или выражения является эффективным признаком, который можно использовать для предсказания типа именованной сущности.

Показано, что можно извлекать эквивалентные варианты именованной сущности. Этот результат способствует решению задачи кластеризации сущностей и семантических отношений моделями, обучающимися без учителя, и может быть использован для поиска перефразировок и автоматического создания онтологий. Модели были обучены на русском сегменте параллельных корпусов, используемых для статистического машинного перевода. Векторные представления были построены и оценены для слов, лексем и словосочетаний.

В работе [22] предлагается гибридная нейросетевая модель распознавания именованных сущностей в медицинских текстах. Предложен метод кодирования, основанный на механизме полного самовнимания (self-attention). Векторное представление каждого слова связывается со всем предложением с помощью механизма внимания. Он определяет распределение веса, оценивая символы или слова во всех позициях, и получает информацию о позиции в предложении, которая требует наибольшего внимания. Вектор кодирования в каждой позиции интегрируется с контекстной информацией полного предложения, что решает проблему неоднозначности (омонимии). Предлагается многомерный сверточный метод декодирования. Этот метод позволяет эффективно учитывать особенности распознавания именованных сущностей в медицинском тексте в процессе декодирования. Он использует двумерное сверточное декодирование для связывания слова в текущей позиции с окружающими словами для повышения эффективности декодирования и извлечения признаков из логики предшествующих и последующих слов. Экспериментальные результаты подтверждают эффективность предложенного метода по сравнению с некоторыми существующими методами распознавания именованных сущностей в медицинских текстах.

В работе [23] применены методы активного обучения для сокращения объема необходимых данных, что позволило за небольшое число итераций повысить эффективность работы моделей глубокого обучения.

С 2018 года для решения задачи извлечения именованных сущностей активно применяются методы глубокого обучения [24] моделей трансформеров (BERT и т.п.), предобученных на большом корпусе неспециализированных текстов, и с помощью подхода на основе переноса обучения (Transfer Learning) с использованием существенно меньшего количества медицинских текстов дообучить модель и существенно повысить ее эффективность в задаче NER.

В работе [7] исследуются различные подходы к распознаванию именованных сущностей в ЭМК. Рассмотрено применение основанных на правилах систем, глубокого обучения и трансферного обучения систем для решения задачи NER в отчетах о визуализации мозга с фокусом на записях пациентов с инсультом. Показано, что наиболее точным способом маркировки ЭМК является экспертная разметка.

В работе [25] использовалась модель, основанная на рекуррентных нейронных сетях (RNN), обученных на большом корпусе не аннотированных медицинских текстов и дообученых с использованием подхода на основе переноса обучения на размеченном наборе документов.

В работе [4] выполнена оценка многозадачного подхода для решения задачи NER на русском языке. Модель основана на комбинации двунаправленной долгой краткосрочной памяти (BiLSTM) и условных случайных полей. Проведены эксперименты на трех существующих русскоязычных аннотированных корпусах.

В статье [5] представлен метод распознавания именованных сущностей NESSMa (Named Entity tagging by Surrounding Sequence Matching), который, основываясь на тегах последовательностей, способен аннотировать слова предложения с помощью нейронной сети с двунаправленной долгой краткосрочной памятью и условным случайным полем (BiLSTM-CRF). В качестве признаков используется вхождение слова в двунаправленном кодирующем представлении, полученном с помощью модели трансформера (BERT), а также теги части речи слова и информация о последовательности меток. Окружение последовательностей каждого слова формируются автоматически из обучающего размеченного набора.

Для оценки эффективности систем NER используются вручную размеченные корпусы текстов. В рамках конференции CoNLL (Conference on Computational Natural Language Learning) предложены оценки точности (precision, P), полноты (recall, R) и F_1-мера:

$$Precision = \frac{\text{число верно выделенных сущностей}}{\text{число всех выделенных сущностей}}, \quad (1)$$

$$Recall = \frac{\text{число верно выделенных сущностей}}{\text{число сущностей в корпусе}}, \quad (2)$$

$$F_1 = \frac{2PR}{P+R}. \quad (3)$$

Для построения корпусов текстов с разметкой именованных сущностей в качестве источника эталонных названий (заболеваний, симптомов, лекарств, процедур и др.) могут быть использованы тезаурусы и онтологии. Например, онтология UMLS (Unified Medical Language System) [26] и тезаурус предметных медицинских рубрик (Medical Subject Headings) [27], который имеет эквивалент на русском языке MSHRUS.

В работе [28] представлен набор данных для нормализации клинических понятий на русском языке RuCCoN (Russian Clinical Concept Normalization), вручную аннотированный медицинскими работниками. Набор данных содержит более 16 028 упоминаний сущностей, вручную связанных с более чем 2 409 уникальными понятиями из русскоязычной части онтологии UMLS.

Проведена оценка качества работы больших языковых моделей на основе предобученных моделей трансформеров с переносом обучения [29]:

1) SapBERT+RuCCoN, с переносом обучения на множестве ЭМК;

2) SapBERT+MCN, с дообучением;

3) SapBERT+WRN [30], с обучением на наборе данных, построенном на основе тезауруса RuWordNet;

4) SapBERT+XL-BEL, на русскоязычной части корпуса XL-BEL;

5) SapBERT+RuCCoN+RWMXL-BEL, комбинация всех трех подходов.

Распознавание именованных сущностей для языков с богатой морфологией является сложной задачей, обусловленной богатством форм отражений и неоднозначности. Эта проблема решается в рамках совместной задачи SlavNER. Предложенная в работе [8] система использует предварительно обученную многоязычную языковую модель BERT и дообучена для шести славянских языков.

Обычные методы NER для медицинских текстов не в полной мере используют немаркированные тексты. В статье [31] представлен подход к распознаванию медицинских именованных сущностей, который включает в себя словарь медицинской терминологии и предварительно обученные языковые модели. Основные усилия направлены на извлечении знаний из немаркированных медицинских текстов.

Создан словарь медицинских именованных сущностей путем анализа размеченных медицинских текстов и сбора размеченных NER с других ресурсов (например, Yidu-N4K [31]). Построенный словарь использован для дообучения предобученных языковых моделей, используя немаркированные медицинские тексты. Применены механизмы разметки немаркированных медицинских текстов для автоматического аннотирования и создания модели тегирования BiLSTM-CRF, которая использовалась для тонкой настройки предварительно обученных языковых моделей. Эксперименты с немаркированными медицинскими текстами, которые были извлечены из электронных медицинских карт, показали, что предложенный подход к NER позволяет достичь оценок F_1 на уровне 88,7% и 95,3% соответственно.

В работе [6] решается задача распознавания именованных сущностей для деидентификации частной и конфиденциальной (PII) из отчетов

независимой медицинской экспертизы (IME). Применен инструментарий NER платформ обработки естественного языка OpenNLP и spaCy. Лучшая модель (spaCy) позволила добиться оценки F_1-средней 0,91 на выборке из 50 отчетов IME из предварительного подготовленного набора данных.

В статье [32] представлен полноразмерный русскоязычный корпус отзывов интернет-пользователей о лекарствах с комплексным распознаванием именованных сущностей с маркировкой фармацевтически значимых объектов. Разметка корпуса включает упоминания следующих сущностей: лекарственный препарат, побочная реакция на лекарственный препарат, болезнь и симптомы. Корпус имеет сложную схему аннотирования, включающую 18 типов упоминаний, пересекающиеся упоминания, прерывистые упоминания и аннотацию кореференции. Мультиклассовая нейросетевая модель для распознавания сущностей, подходящая для маркировки представленного корпуса, разработана на основе объединения языковой модели XLM-RoBERTa с выбранным набором входных признаков.

В [33] выделяется несколько инструментов для выявления именованных сущностей, работающие с русским языком:

– DeepPavlov BERT NER: SOTA-система для русского языка – имеет возможность обучения;

– slovnet BERT NER: DeepPavlov BERT NER с дистилляцией моделей;

– синтетическая разметка (Nerus) в WordCNN-CRF с квантованными векторами вложений (Navec);

– spaCy: модель tok2vec и Multilingual BERT.

Таким образом, на сегодняшний день наиболее актуальным является применение больших языковых моделей в сочетании с технологиями переноса обучения на специализированных корпусах текстов для построения предметно-ориентированных систем распознавания именованных сущностей.

3. Эксперимент по построению системы NER на основе модели-трансформера с переносом обучения

В [34] показано, что языковые модели общего назначения, не дообученные на размеченном корпусе, не позволяют приблизиться по качеству выделения именованных сущностей к специализированным моделям, однако, позволяют предварительно разметить корпус для дальнейшей верификации и уточнения разметки.

Для эксперимента по построению системы NER выбрана модель трансформер DeepPavlov BERT NER в составе фреймворка spaCy. Далее модель была дообучения с помощью подхода на основе переноса обучения на специализированном корпусе текстов. Исходный корпус текстов проблемной области детально описан в [35, 36] и включает 162 документа обезличенных историй болезни пациентов педиатрического центра с болезнями органов дыхания, с аллергическими, нефрологическими и ревматическими болезнями.

Разработанный модуль NER состоит из этапов обработки, представленных на рисунке 1 [37].

Для тестирования модуля были сделаны следующие шаги:

– предварительная структуризация данных;
– передача данных для обработки;
– запуск обработки данных и проверка результата.

На рисунке 2 видно, что модуль отработал в штатном режиме, разметив все сущности, относящие к симптомам и названиям болезни.

В таблице 2 показаны основные метрики качества на тестовой выборке для построенного решения после реализации переноса обучения.

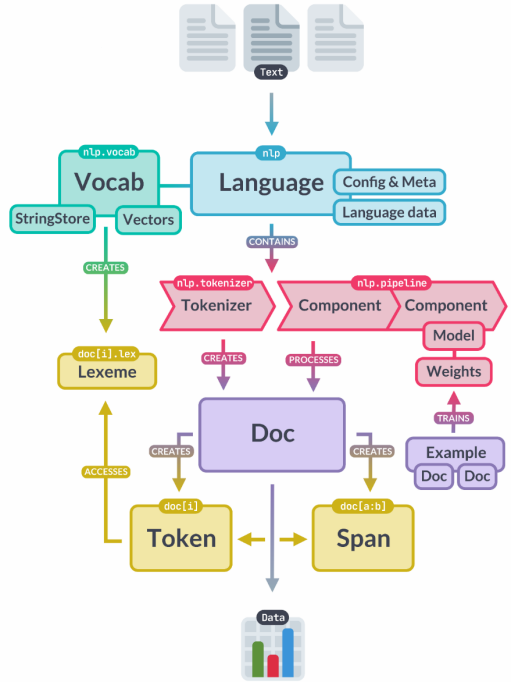

Рис. 1. Схема взаимодействия модулей программного обеспечения системы NER

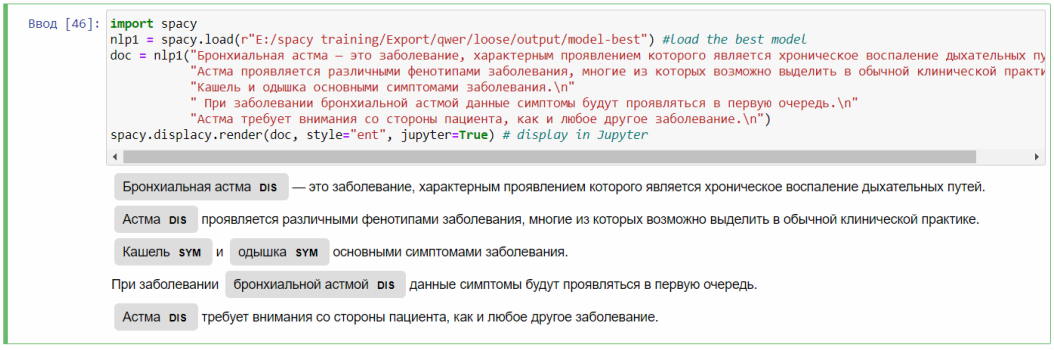

Рис. 2. Пример выделения именованных сущностей

Табл. 2. Основные метрики качества работы системы NER на тестовой выборке

Параметр	Базовая модель	Дообученная модель
Точность (P)	67 %	74 %
Полнота (R)	64 %	71 %
F_1-мера	65 %	72 %

Эксперимент на корпусе медицинских текстов заданной тематики показал, что модели общего назначения не позволяют приблизиться по качеству выделения именованных сущностей к экспертной разметке. Однако перенос обучения позволяет существенно повысить эффективность моделей даже на небольших (относительно количества параметров модели) размеченных текстах.

Заключение

Применение языковых моделей трансформеров для выделения именованных сущностей позволяют расширить возможности по формализации знаний и решению задач построения систем поддержки принятия решений в клинической практике. Необходимым этапом является детализированная разметка корпуса текстовых документов для учета специфической лексики и обилия сокращений для последующего дообучения глубоких нейросетевых моделей. Перенос обучения позволяет существенно повысить эффективность моделей (на 5-8 %) даже на небольших (относительно количества параметров модели) размеченных корпусах текстов.

Финансирование. Работа выполнена при финансовой поддержке РНФ (проект № 22-19-00471).

Список литературы

1. Ижунинов М.А. Big Data в здравоохранении // Молодой ученый. – 2019. – № 50(288). – С. 8-10.
2. Zhukova N., Berezov M., Lebedev S., Zavadskaya E. Extraction of Named Entities from Semi-Structured Texts for Medical Domain // AIST (Supplement). 2017, pp. 31-40.
3. Дудченко П.В. Проблемы извлечения данных из медицинских выписок // Наука, техника и образование. – 2018. – № 10 (51). – С. 36-38.
4. Mazitov D., Alimova I., Tutubalina E. Named entity recognition in Russian using multi-task LSTM-CRF // Journal of Mathematical Sciences. 2023, vol. 273, № 4, pp. 1-10.
5. Landolsi M.Y., Romdhane L.B., Hlaoua L. Medical named entity recognition using surrounding sequences matching // Procedia Computer Science. 2022, vol. 207, pp. 674-683.
6. Извлечение/распознавание сущностей для обучения моделей НЛП [Электронный ресурс]. – Режим доступа: https://ru.shaip.com/healthcare-ai/clinical-ner/.
7. Pearson C., Seliya N., Dave R. Named entity recognition in unstructured medical text documents // 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET). 2021, pp. 1-6.
8. Viksna R., Skadiņa I. Multilingual slavic named entity recognition // Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. – Kiev: Association for Computational Linguistics, 2021. – P. 93-97.
9. Киреев Д.А. Интеллектуальный анализ клинических текстов // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: Материалы Всероссийской конференции с международным участием. – М.: РУДН, 2021. – С. 209-213.

10. 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.
11. NLP. Основы. Техники. Саморазвитие. Часть 2: NER [Электронный ресурс]. – Режим доступа: https://habr.com/ru/companies/contentai/articles/449514/.
12. Wu M., Liu F., Cohn T. Evaluating the Utility of Hand-crafted Features in Sequence Labelling // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – Brussels: Association for Computational Linguistics, 2018. – P. 2850-2856.
13. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. Neural Architectures for Named Entity Recognition // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – San Diego: Association for Computational Linguistics, 2016. – P. 260-270.
14. Neamatullah I., Douglass M.M., Lehman L.W.H., Reisner A., Villarroel M., Long W.J., Clifford G.D. Automated de-identification of free-text medical records // BMC medical informatics and decision making. 2008, vol. 8, no. 1, pp. 1-17.
15. Донитова В.В., Киреев Д.А., Титова Е.В., Акимова А.А. Методы обработки естественного языка для извлечения факторов риска инсульта из медицинских текстов // Труды ИСА РАН. – 2021. – Т. 71, № 4. – С. 93.
16. Sondhi P., Gupta M., Zhai C., Hockenmaier J. Shallow Information Extraction from Medical Forum Data // Coling 2010: Posters. 2010, pp. 1158-1166.
17. Nayel H., Shashirekha H.L. Improving {NER} for Clinical Texts by Ensemble Approach using Segment Representations // Proceedings of the 14th International Conference on Natural Language Processing. – Kolkata: NLPAI, 2017. – P. 197-204.
18. Arbabi A., Adams D.R., Fidler S., Brudno M. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning // JMIR medical informatics. 2019, vol. 7, № 2, pp. e12596.
19. Gavrilov D., Gusev A., Korsakov I., Novitsky R., Serova L. Feature extraction method from electronic health records in Russia // Conference of Open Innovations Association, FRUCT. – FRUCT Oy, 2020. – № 26. – P. 497-500.
20. Hahn U., Oleynik M. Medical Information Extraction in the Age of Deep Learning // Yearbook of medical informatics. 2020, vol. 29, no 01, pp. 208-220.
21. Ivanitskiy R., Shipilo A., Kovriguina L. Russian Named Entities Recognition and Classification Using Distributed Word and Phrase Representations // SIMBig. 2016, pp. 150-156.
22. Yang T., He Y., Yang N. Named entity recognition of medical text based on the deep neural network // Journal of Healthcare Engineering. 2022, vol. 2022, pp. 1-10.
23. Shelmanov A., Liventsev V., Kireev D., Khromov N., Panchenko A., Fedulova I., Dylov D.V. 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, pp. 482-489.
24. Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics. 2020, vol. 36, no. 4, pp. 1234-1240.

25. Gligic L., Kormilitzin A., Goldberg P., Nevado-Holgado A. Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks // Neural Networks. 2020, vol. 121, pp. 132-139.
26. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology // Nucleic acids research. 2004, vol. 32, no. suppl_1, pp. D267-D270.
27. Coletti M.H., Bleich H.L. Medical subject headings used to search the biomedical literature // Journal of the American Medical Informatics Association. 2001, vol. 8, no. 4, pp. 317-323.
28. Nesterov A., Zubkova G., Miftahutdinov Z., Kokh V., Tutubalina E., Shelmanov A., Nikolenko S. RuCCoN: clinical concept normalization in Russian // Findings of the Association for Computational Linguistics: ACL 2022. – Dublin: Association for Computational Linguistics, 2022. – P. 239-245.
29. Sung M., Jeon H., Lee J., Kang J. Biomedical entity representations with synonym marginalization // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, pp. 3641-3650.
30. Fair-Evaluation-BERT [Electronic resource]. – Access mode: https://github.com/insilicomedicine/Fair-Evaluation-BERT.
31. Wen C., Chen T., Jia X., Zhu J. Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary // Data Intelligence. 2021, vol. 3, no. 3, pp. 402-417.
32. Sboev A., Sboeva S., Moloshnikov I., Gryaznov A., Rybka R., Naumov A., Ilyin V. Analysis of the full-size Russian corpus of internet drug reviews with complex NER labeling using deep learning neural networks and language models // Applied Sciences. 2022, vol. 12, no. 1, pp. 491.
33. Соколовский Д.Е., Землянский С.А. Использование инструмента DeepPavlov для извлечения и структурирования собственных именованных сущностей из медицинских наборов данных // Молодежь и современные информационные технологии: сборник трудов XIX Международной научно-практической конференции студентов, аспирантов и молодых учёных. – Томск: ТПУ, 2022. – С. 238-239.
34. Зулкарнеев Р.Х., Юсупова Н.И., Сметанина О.Н., Гаянова М.М., Вульфин А.М. Методы и модели извлечения знаний из медицинских документов // Информатика и автоматизация. – 2022. – Т. 21, № 6. – С. 1169-1210.
35. Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В., Девяткин Д.А., Шелманов А.О., Вишнёва Е.А., Смирнов В.И. Технологии комплексного интеллектуального анализа клинических данных // Вестник Российской академии медицинских наук. – 2016. – Т. 71, № 2. – С. 160-171.
36. Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В., Девяткин Д.А., Шелманов А.О., Вишнева Е.А., Латышев А.В. Методы и средства комплексного интеллектуального анализа медицинских данных // Труды ИСА РАН. – 2015. – Т. 65, № 2. – С. 81-93.
37. Industrial-Strength Natural Language Processing [Electronic resource]. – Access mode: https://spacy.io/.

References

1. Izhuninov M.A. Big Data in healthcare // Young scientist. 2019, no. 50 (288), pp. 8-10.

2. Zhukova N., Berezov M., Lebedev S., Zavadskaya E. Extraction of Named Entities from Semi-Structured Texts for Medical Domain // AIST (Supplement). 2017, pp. 31-40.
3. Dudchenko P.V. Problems of extracting data from medical records // Science, technology and education. 2018, no. 10 (51), pp. 36-38.
4. Mazitov D., Alimova I., Tutubalina E. Named entity recognition in Russian using multi-task LSTM-CRF // Journal of Mathematical Sciences. 2023, vol. 273, no. 4, pp. 1-10.
5. Landolsi M.Y., Romdhane L.B., Hlaoua L. Medical named entity recognition using surrounding sequences matching // Procedia Computer Science. 2022, vol. 207, pp. 674-683.
6. Entity extraction/recognition for training NLP models [Electronic resource]. – Access mode: https://ru.shaip.com/healthcare-ai/clinical-ner/.
7. Pearson C., Seliya N., Dave R. Named entity recognition in unstructured medical text documents // 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET). 2021, pp. 1-6.
8. Viksna R., Skadiņa I. Multilingual slavic named entity recognition // Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. – Kiev: Association for Computational Linguistics, 2021. – P. 93-97.
9. Kireev D.A. Intellectual analysis of clinical texts // Information and telecommunication technologies and mathematical modeling of high-tech systems: Materials of the All-Russian conference with international participation. – M.: RUDN, 2021. – P. 209-213.
10. Serdyuk Yu.P., Vlasova N.A., Momot S.R. System for extracting symptom references from natural language texts using neural networks // Software systems: theory and applications. 2023, vol. 14, no. 1 (56), pp. 95-123.
11. NLP. Basics. Technicians. Self-development. Part 2: NER [Electronic resource]. – Access mode: https://habr.com/ru/companies/contentai/articles/449514/.
12. Wu M., Liu F., Cohn T. Evaluating the Utility of Hand-crafted Features in Sequence Labelling // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. – Brussels: Association for Computational Linguistics, 2018. – P. 2850-2856.
13. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. Neural Architectures for Named Entity Recognition // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – San Diego: Association for Computational Linguistics, 2016. – P. 260-270.
14. Neamatullah I., Douglass M.M., Lehman L.W.H., Reisner A., Villarroel M., Long W.J., Clifford G.D. Automated de-identification of free-text medical records // BMC medical informatics and decision making. 2008, vol. 8, no. 1, pp. 1-17.
15. Donitova V.V., Kireev D.A., Titova E.V., Akimova A.A. Natural language processing methods for extracting risk factors for stroke from medical texts // Proceedings of ISA RAS. 2021, vol. 71, no. 4, pp. 93.
16. Sondhi P., Gupta M., Zhai C., Hockenmaier J. Shallow Information Extraction from Medical Forum Data // Coling 2010: Posters. 2010, pp. 1158-1166.
17. Nayel H., Shashirekha H.L. Improving {NER} for Clinical Texts by Ensemble Approach using Segment Representations // Proceedings of the 14th International Conference on Natural Language Processing. – Kolkata: NLPAI, 2017. – P. 197-204.

18. Arbabi A., Adams D.R., Fidler S., Brudno M. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning // JMIR medical informatics. 2019, vol. 7, № 2, pp. e12596.
19. Gavrilov D., Gusev A., Korsakov I., Novitsky R., Serova L. Feature extraction method from electronic health records in Russia // Conference of Open Innovations Association, FRUCT. – FRUCT Oy, 2020. – No 26. – P. 497-500.
20. Hahn U., Oleynik M. Medical Information Extraction in the Age of Deep Learning // Yearbook of medical informatics. 2020, vol. 29, no 01, pp. 208-220.
21. Ivanitskiy R., Shipilo A., Kovriguina L. Russian Named Entities Recognition and Classification Using Distributed Word and Phrase Representations // SIMBig. 2016, pp. 150-156.
22. Yang T., He Y., Yang N. Named entity recognition of medical text based on the deep neural network // Journal of Healthcare Engineering. 2022, vol. 2022, pp. 1-10.
23. Shelmanov A., Liventsev V., Kireev D., Khromov N., Panchenko A., Fedulova I., Dylov D.V. 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, pp. 482-489.
24. Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics. 2020, vol. 36, no. 4, pp. 1234-1240.
25. Gligic L., Kormilitzin A., Goldberg P., Nevado-Holgado A. Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks // Neural Networks. 2020, vol. 121, pp. 132-139.
26. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology // Nucleic acids research, 2004, vol. 32, no. suppl_1, pp. D267-D270.
27. Coletti M.H., Bleich H.L. Medical subject headings used to search the biomedical literature // Journal of the American Medical Informatics Association. 2001, vol. 8, no. 4, pp. 317-323.
28. Nesterov A., Zubkova G., Miftahutdinov Z., Kokh V., Tutubalina E., Shelmanov A., Nikolenko S. RuCCoN: clinical concept normalization in Russian // Findings of the Association for Computational Linguistics: ACL 2022. – Dublin: Association for Computational Linguistics, 2022. – P. 239-245.
29. Sung M., Jeon H., Lee J., Kang J. Biomedical entity representations with synonym marginalization // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, pp. 3641-3650.
30. Fair-Evaluation-BERT [Electronic resource]. – Access mode: https://github.com/insilicomedicine/Fair-Evaluation-BERT.
31. Wen C., Chen T., Jia X., Zhu J. Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary // Data Intelligence. 2021, vol. 3, no. 3, pp. 402-417.
32. Sboev A., Sboeva S., Moloshnikov I., Gryaznov A., Rybka R., Naumov A., Ilyin V. Analysis of the full-size Russian corpus of internet drug reviews with complex NER labeling using deep learning neural networks and language models // Applied Sciences. 2022, vol. 12, no. 1, pp. 491.
33. Sokolovsky D.E., Zemlyansky S.A. Using the DeepPavlov tool to extract and structure your own named entities from medical datasets // Youth and modern information technologies: collection of proceedings of the XIX International Scientific and

Practical Conference of Students, Postgraduate Students and Young Scientists. – Tomsk: TPU, 2022. – P. 238-239.
34. Zulkarneev R.Kh., Yusupova N.I., Smetanina O.N., Gayanova M.M., Vulfin A.M. Methods and models for extracting knowledge from medical documents // Computer Science and Automation. 2022, vol. 21, no. 6, pp. 1169-1210.
35. Baranov A.A., Namazova-Baranova L.S., Smirnov I.V., Devyatkin D.A., Shelmanov A.O., Vishneva E.A., Smirnov V.I. Technologies for complex intellectual analysis of clinical data // Bulletin of the Russian Academy of Medical Sciences. 2016, vol. 71, no. 2, pp. 160-171.
36. Baranov A.A., Namazova-Baranova L.S., Smirnov I.V., Devyatkin D.A., Shelmanov A.O., Vishneva E.A., Latyshev A.V. Methods and tools for complex intellectual analysis of medical data // Proceedings of ISA RAS. 2015, vol. 65, no. 2, pp. 81-93.
37. Industrial-Strength Natural Language Processing [Electronic resource]. – Access mode: https://spacy.io/.

Вульфин Алексей Михайлович – доктор технических наук, профессор	**Vulfin Alexey Mikhailovich** – doctor of technical sciences, professor
Гаянова Майя Марсовна – кандидат технических наук, доцент	**Gayanova Maya Marsovna** – candidate of technical sciences, associate professor
Улямаев Тимур Ильшатович – бакалавр	**Ulyamaev Timur Ilshatovich** – bachelor
vulfin.alexey@gmail.com	

Received 09.12.2023