

ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТОВ MEDSPACy И DEERPAVLOV ДЛЯ ИЗВЛЕЧЕНИЯ СИМПТОМОВ COVID-19 И ЛЕКАРСТВ ИЗ МЕДИЦИНСКИХ ТЕКСТОВ

Соколовский Д.Е.

Томский политехнический университет, ИШИТР, аспирант гр. А1-39,

e-mail: des16@tpu.ru

Аннотация

Протестированы инструменты MedSpaCy и DeepPavlov для извлечения именованных сущностей (симптомов заболеваний и медикаментов с дозировками соответственно), которые показали значительное улучшение точности в этой части системы. По результатам анализа статей исследователей, экспериментов и тестирования работы инструментов, применен и протестирован метод гибридного подхода для увеличения точности системы по F1-мере.

Ключевые слова: извлечение именованных сущностей, BERT, SpaCy, MedSpaCy, DeepPavlov.

Введение

Одной из важных причин роста популярности инструментов и технологий машинного обучения является накопление больших массивов неструктурированных данных во многих аспектах деятельности человека. Здравоохранение и медицина относятся к таким отраслям, которые генерируют значительный объем разнородных данных (результаты многочисленных анализов в бумажном и электронном виде, ЭКГ, КТ, дневники самонаблюдений, осмотры специалистом и т.д.) [1].

Медицинские данные нужно обрабатывать, применяя обособленные или комбинируя инструменты и нейросети для достижения приемлемых результатов. Одним из способом структурирования разрозненного текста является извлечение именованных сущностей из текста с последующим их структурированием и составлением основной информации. Именованная сущность при извлечении информации - это объект реального мира, такой как имя человека, локация, названия организаций и т.д., который может быть обозначен собственным именем. Она может быть абстрактной или иметь физическое существование, например для медицинских данных - именованной сущностью может являться само слово температура, артериальное давление и их значения, а также симптомы или другие признаки заболевания [2].

В настоящее время для извлечения именованных сущностей и обработки естественного языка на языке Python по статистике зачастую используют один из самых популярных инструментов Spacy. Функционал инструмента позволяет решать очень широкий спектр задач: от определения частей речи и выделения именованных сущностей до создания собственных моделей для анализа, в том числе и медицинских данных [3-4].

Описание и тестирование алгоритма

Проведя исследование двух лучших на данный момент инструментов по извлечению именованных сущностей Spacy и DeepPavlov, а также проведя тестирование на практических данных, было выявлено, что работа BERT обеих моделей показала близкие результаты с небольшим преимуществом DeepPavlov, 0.91 против 0.93 по F1-мере соответственно, но клиентская часть, визуализация и дополнительные инструменты экосистемы SpaCy на примере сущностей температуры (Темп), артериального давления (АД), частоты сердечных сокращений (ЧСС), частоты дыхательных движений (ЧДД) и их значений, работают во многих случаях быстрее и удобнее [5], после чего сделан ошибочный вывод, что этот инструмент можно применять обособленно. Корректный результат отработки обособленной модели SpaCy представлен на рис. 1.

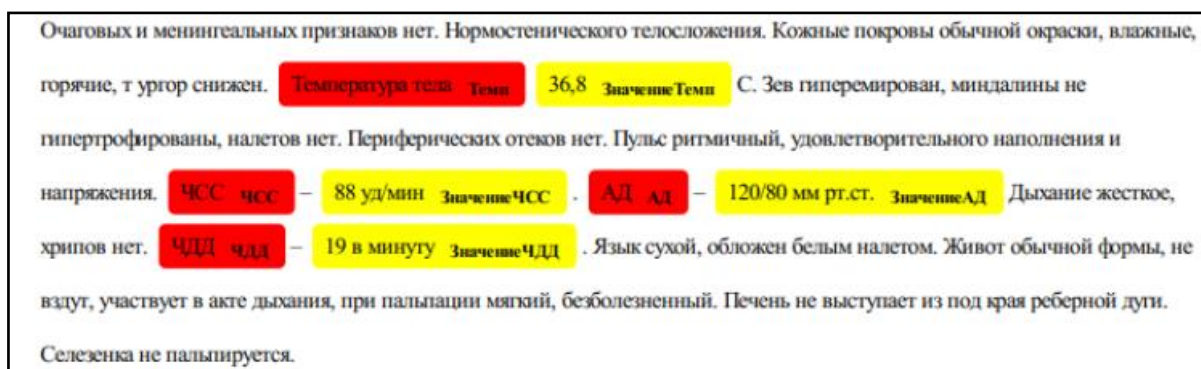


Рис. 1. Распознавание численных сущностей COVID-19

Так как протестировав обособленный подход на выявление симптомов заболеваний и медикаментов, нейросеть не может в полной мере охватить все возможные названия и значения сущностей, если при тренировке были использованы ограниченные наборы данных. В таком случае, она ориентируется только на те данные, на которых обучалась, но этого недостаточно для полноценной работы и корректного выявления сущностей.

Изучив статьи исследователей по данной тематике о целесообразности применения гибридных подходов, был сделан вывод, что к инструменту Sрасу нужно добавить другие инструменты, подходы и методы, чтобы каждый из них выполнял только нужную часть работы, разгружая систему своей модульностью [6]. В результате недостатки, которые получены при тестировании обособленного подхода по выявлению симптомов заболеваний, можно исключить при использовании еще одного действенного метода по извлечению именованных сущностей - регулярных выражений [7]. Регулярные выражения — формальный язык, используемый в компьютерных программах, работающих с текстом, для поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов. Для поиска используется строка-образец, состоящая из символов и метасимволов и задающая правило поиска. Несмотря на то, что регулярные выражения после тестирования достаточно точно (в подавляющем большинстве случаев близко к 100 %) определяет именованные сущности и их значения, они тоже не могут работать полноценно и обособленно, так как невозможно все данные из исходного текста предугадать корректно при построении регулярного выражения. Исходя из анализа результатов, принято решение объединить работу нейросетей, регулярных выражений для основных именованных сущностей, а для выявления симптомов заболеваний или лекарственных препаратов и их дозировок применять иные инструменты, которые закроют все найденные недостатки, для полноценной и обособленной работы системы по извлечению именованных сущностей.

Для такой задачи протестированы специализирующийся на анализе медицинских данных MedSpaCy [8], работающий в экосистеме Sрасу, а также уже проверенный инструмент DeepPavlov [9]. По результатам обучения и тестирования инструмент MedSpaCy показал себя удобным в использовании извлечения именованных сущностей медикаментов, так как имеет в своем составе возможность использования предварительной обработки исходного текста, например приведение к нижнему регистру, точечных регулярных выражений с использованием функции вариации элемента, например, если медикамент содержит в названии цифры, которые могут быть записаны по-разному (2, II), либо названия препаратов имеющих сокращения (Артериальное давление, АД). Модель получила оценку F1-меры 0.92, целесообразно применять MedSpaCy в части дневника пациента “План лечения”, где имеются медикаменты и дозировки, результат корректного применения представлен на рис. 2. Варианты с запятой данный инструмент воспринимает плохо, поэтому текст был приведен к английской системе чисел с точкой с помощью функции locale.

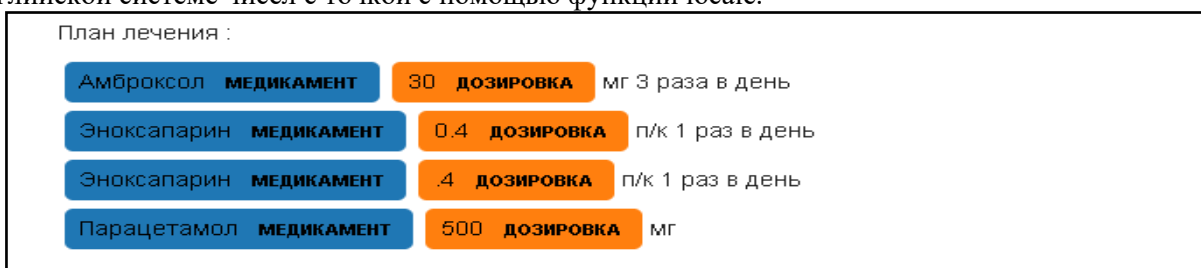


Рис. 2. Распознавание медикаментов и дозировок

Обучив модель DeepPavlov и протестировав ее на практических медицинских данных выявления симптомов, модель получила оценку по F1-мере 0.95. Целесообразно применение модели только для выявления симптомов, результат корректного применения на рис. 3.

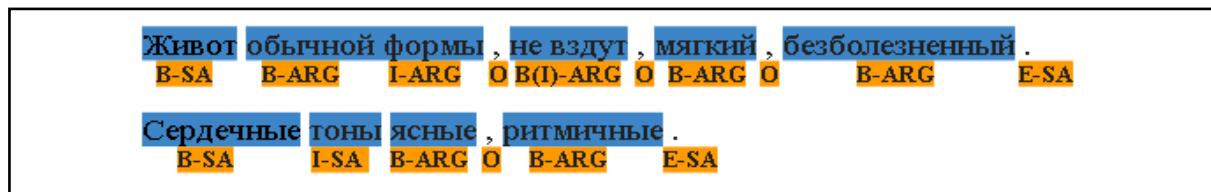


Рис. 3. Распознавание симптомов

Здесь метки b-sa (begin-symptom with arguments) и e-sa (end-symptom with arguments) обрамляют всё выражение с перечислением симптомов, i-sa (inside-symptom with arguments) отмечают отдельное слово внутри симптома, а метки b-arg (begin-argument), а также i-arg (inside-argument) отмечают отдельные аргументы.

Заключение

Протестированы инструменты MedSpaCy и DeepPavlov для извлечения именованных сущностей (симптомов заболеваний и медикаментов с дозировками соответственно), которые показали значительное улучшение точности в этой части системы. По результатам анализа статей исследователей, экспериментов и тестирования работы инструментов по выявлению именованных сущностей развитие системы и дальнейшее тестирование отмечено, как содержащее в своей основе схему гибридного подхода, в том числе с инструментами, представленными в текущей статье для увеличения общей точности по F1-мере.

Список использованных источников

1. Карнаухов Н.С., Ильяхин Р.Г. Возможности технологий «Big Data» в медицине // Врачи и информационные технологии. – 2019. – No 1. – С. 59–63.
2. Taufiq, U. Named entity recognition and dependency parsing for better concept extraction in summary obfuscation detection / U. Taufiq, R. Pulungan, Y. Suyanto // Expert Systems with Applications. – 2023. – Vol. 217. – P. 119579. – DOI 10.1016/j.eswa.2023.119579.
3. Fabienne Krauer, Boris V. Schmid. Mapping the plague through natural language processing // Epidemics 41. – 2022.
4. Spacy [Электронный ресурс]. – URL: <https://spacy.io> (дата обращения: 15.01.2024).
5. Соколовский Д.Е., Некрасов В.Н., Землянский С.А., Аксенов С.В. Оценка использования инструментов библиотеки SpaCy и DeepPavlov для задачи извлечения именованных сущностей из описаний результатов осмотров пациентов с COVID-19 // Известия Томского политехнического университета. Промышленная кибернетика – 2023. – Т. 1. – No 2. 46–53
6. Daniel X Le, James G Mork, Sameer Antani.: Hybrid Ensemble-Rule Algorithm for Improved MEDLINE® Sentence Boundary Detection. AMIA Annu Symp Proc. – Vol. 2021. – P. 677-686.
7. Sander Puts, Martijn Nobel, Catharina Zegers, Iñigo Bermejo, Simon Robben, Andre Dekker.: How Natural Language Processing Can Aid With Pulmonary Oncology Tumor Node Metastasis Staging From Free-Text Radiology Reports: Algorithm Development and Validation. JMIR Form Res. – Vol. 7. – P. 38125
8. MedSpaCy [Электронный ресурс]. – URL: <https://spacy.io/universe/project/medspacy/> (дата обращения: 23.01.2024)
9. DeepPavlov [Электронный ресурс]. – URL: <https://github.com/deepmipt/DeepPavlov> (дата обращения: 23.01.2024)