ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТА SPACY ДЛЯ ИЗВЛЕЧЕНИЯ ЧАСТНЫХ ИМЕНОВАННЫХ СУЩНОСТЕЙ COVID-19 ИЗ МЕДИЦИНСКИХ НАБОРОВ ДАННЫХ

Соколовский Д.Е.
НИ ТПУ, ИШИТР, А1-39, e-mail: des16@tpu.ru

Введение

В работе рассматривается возможности инструмента для выявления именованных сущностей «Spacy». Для его анализа было проведено обучение своей модели на основе существующей, а также тестирование работы модели на медицинских наборах данных (дневниках пациентов) для их дальнейшего структурирования [1].

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 [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].

Описание и тестирование алгоритма

В работе был рассмотрен процесс обучения модели с помощью инструмента Spacy (версия 3.3), а также процесс тестирования и доработки модели для точного определения частных именованных сущностей. Для начала работы с библиотекой Spacy необходимо выбрать на официальном сайте уже обученную модель языка для работы с текстом, которая будет являться основой и дополнить ее своими сущностями. В нашем случае выбрана за основу ru_core_news_lg, т.к. мы обучаем модель на русскоязычных текстовых данных, и она уже содержит некоторые сущности, которые могут быть использованы при выявлении [4]. Также для перехода к обучению модели был сформирован тренировочный конфиг для выявления NER необходимый для корректной работы инструмента. Для его создания на официальном сайте инструмента Spacy в разделе Training models Quickstart нужно выбрать русский язык, компонент «NER» и сохранить сгенерированный код [5].

Для обучения модели по выявлению медицинских именованных сущностей, были подготовлены и размечены 3 документа в формате pdf (trainpdf, trainpdf2, trainpdf3). Эти файлы представляют собой первичный осмотр врача в приёмном отделении и связаны с Covid-19. В них были выделены красным цветом показатели состояния больного (Температура тела, Частота дыхательных движений (ЧДД), Частота сердечных сокращений (ЧСС), Артериальное давление (АД)), а желтым значения показателей. После подготовки тренировочных файлов и конфига, была написана программа для повышения скорости и удобства тренировки модели распознавания медицинских именованных сущностей [6]. Для более точного обучения модели в коде программы, в переменную «TRAIN_DATA», необходимо сформировать и поместить как минимум один дополнительный вариант тренировочных данных, основывающиеся на дневниках пациентов иначе при дальнейшем обучении могут возникнуть неточности с определением сущностей. Шаблон таких данных выглядит следующим образом «("Текст", {'entities': [(начальный индекс элемента, конечный индекс элемента, 'Тег')]}).»

Переменные, использованные в данном шаблоне:

- текст (любой текст, любого размера, который мы используем для обучения);
- начальный индекс элемента (начальный индекс слова, которое мы используем для маркировки);
- конечный индекс элемента (индекс элемента, следующий после последнего индекса слова, которое мы используем для маркировки);
- тег (название найденной именованной сущности).

Процесс обучения также можно контролировать в терминале и при успешном обучении появится оповещение о пути сохранения модели.

После обучения, модель была протестирована на новых и неразмеченных дневниках пациентов, в количестве 14 штук. На рисунке 1 представлена часть файла, где модель произвела разметку и указала

название показателя и название значения показателя, файл «115-pages» распознался моделью полностью успешно.

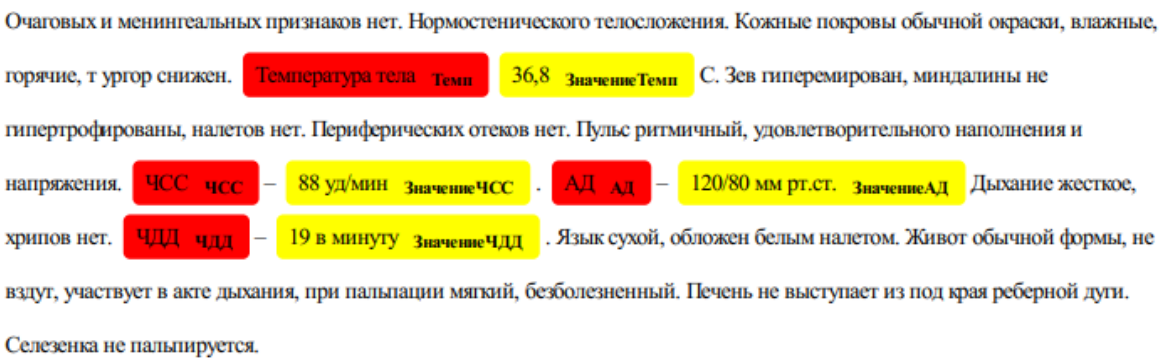

Рис. 1. Распознавание именованных сущностей

Модель работает успешно, но в некоторых файлах не определялось значение температуры. Проанализировав исходный текст, было выявлено то, что в таких файлах значение температуры указано слитно с единицей измерения, например «36.6С». Доработав программный код, который отделяет значение от единицы измерения, модель стала точно определять все сущности, указанные при тренировке, во всех тестовых файлах.

Заключение

По результатам экспериментов исследования и тестирования работы Spacy (версия 3.3), который имеет в своем функционале возможность обучения собственных моделей на своем наборе данных, обучена собственная модель и протестирована на медицинских данных. Модель на тестовом этапе без доработок имеет показатель f1-меры 0,80. Протестирован подход по доработке исходных данных с помощью библиотек обработки естественного языка, который увеличил точность модели.

Список использованных источников

1. Spacy [Электронный ресурс]. – URL: https://spacy.io (дата обращения: 01.02.2023).
2. Umar Taufiq, Reza Pulungan, Yohanes Suyanto.: Named entity recognition and dependency parsing for better concept extraction in summary obfuscation detection. Expert Systems with Applications 2017, (2023).
3. Fabienne Krauer, Boris V. Schmid.: Mapping the plague through natural language processing. Epidemics 41, (2022).
4. Spacy Models [Электронный ресурс]. – URL: https://spacy.io/models/ru (дата обращения: 01.02.2023).
5. Training Pipelines & Models [Электронный ресурс]. – URL: https://spacy.io/usage/training#quickstart (дата обращения: 01.02.2023).
6. Seyede Faezeh Mousavi, Mohammadamin Ebrahimi, Seyed Amirhosein Ahmadpour Moghaddam, Narges Moafi, Mahbobe Jafari, Ayoub Tavakolian, Mohsen Heidary.: Evaluating the characteristics of patients with SARS-CoV-2 infection admitted during COVID-19 peaks: A single-center study. Vacunas 24, 27–36 (2023).