ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТА DEEPPAVLOV ДЛЯ ИЗВЛЕЧЕНИЯ И СТРУКТУРИРОВАНИЯ СОБСТВЕННЫХ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ МЕДИЦИНСКИХ НАБОРОВ ДАННЫХ

Д.Е. Соколовский, аспирант
Томский политехнический университет
С.А. Землянский, аспирант
Томский государственный университет
E-mail:des16@tpu.ru

Введение

В работе рассматриваются инструменты для выявления именованных сущностей, обучение своей модели на существующей, а также тестирование работы модели на медицинских наборах данных (дневниках пациентов) для их дальнейшего структурирования.

Именованная сущность — это слово или словосочетание, обозначающее предмет или явления определенной категории. В понятие именованных сущностей входят имена людей, названия организаций, локаций и другие [1].

Пример извлечения именованных сущностей представлен на рисунке 1.

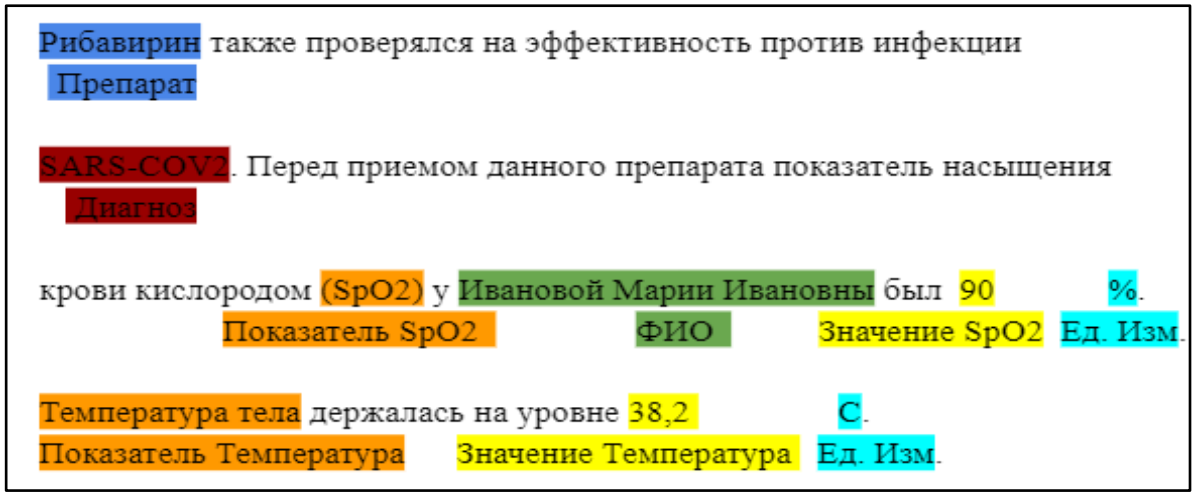

Рис. 1. Извлечение именованных сущностей

Выделяется несколько инструментов для выявления именованных сущностей, работающие с русским языком и python:

- DeepPavlov BERT NER: SOTA-система (наилучшие на данный момент результаты) для русского языка – имеет возможность обучения [2];
- slovnet BERT NER: аналог DeepPavlov BERT NER + дистилляция
- через синтетическую разметку (Nerus) в WordCNN-CRF с квантованными эмбеддингами (Navec) + движок для инференса на NumPy [3];
- spaCy: предлагает tok2vec и Multilingual BERT. Также, пользователь может самостоятельно выбрать предобученную модель из списка на HuggingFace [4].

Описание и тестирование алгоритма

В данной работе нам был рассмотрен инструмент Deeppavlov BERT (версия 0.17.2) и протестирован на практике [5].

С помощью него будут извлекаться частные сущности и их значения, созданной с помощью машинного обучения моделью, которая обучается на примере дневников пациентов, а также выявляться даты и имена для структурированного и быстрого получения медицинских данных.

Разметка для обучения производилась путем определения группы каждой сущности “Температура тела” определена как TEMP, значение температуры, например 37,2, TEPMVALUE, а единица измерения, например “градусы”, TEMPMETR. “Артериальное давление” определено

аналогично AD, значение, например 120/80, ADVALUE, а единица измерения, например “мм. рт. ст.”, ADMETR), а также дата, как (DATE), а частота дыхательных движений, как CHDD. Весь остальной неразмеченный текст помечается «O».

Для тестового анализа были подготовлены 3 файла (train.txt, valid.txt, test.txt), в виде размеченного текста.

Для работы с библиотекой DeepPavlov необходимо выбрать модель языка для работы с текстом. Была выбрана ner_ontonotes_bert_mult, т.к. она является мультиязычной и содержит не только русский язык, но в том числе имеется поддержка английского языка, что очень важно для обработки медицинских данных, которые могут иметь не только термины русского языка [6]. После этого разрабатывается собственная модель для выявления именованных сущностей на основе уже имеющийся, которая обучается на подготовленных данных и на тестовом этапе показывает f1 меру на уровне около 75% (в качестве ошибок, были, например следующие, неправильно определена именованная сущность частота дыхательных движений (ЧДД), вместо нее определилась “ЧСС”).

На первом этапе мы получаем числовые значения и через них получить информацию к какой сущности относится данное значение. Но изначально для того, чтобы поместить в модель нужные нам данные, мы проводим автоматизированную предобработку оригинального текста и помещаем все нужные нам слова и значения в кавычки.

А уже после этого помещаем в модель и получаем значения именованных сущностей, таких как (ADVALUE и TEMPVALUE).

Получая эти данные, мы формируем json файл, структурируя оригинальный дневник пациента, который представлен на рисунке 2.

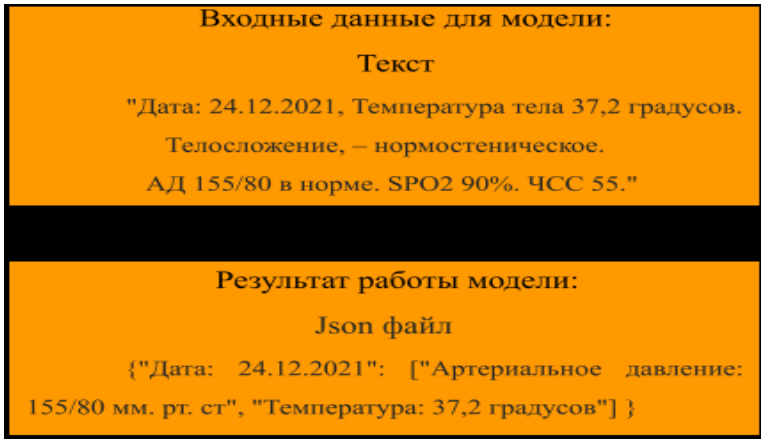

Рис. 2. Процесс формирования json файла

Заключение

По результатам экспериментов были рассмотрены популярные инструменты по выявлению именованных сущностей, выбран для исследования и тестирования работы Deeppavlov BERT (версия 0.17.2), который имеет в своем функционале возможность обучения собственных моделей на своем наборе данных, обучена собственная модель и протестирована на медицинском тексте. Данная модель на тестовом этапе имеет показатель f1-меры 0,75.

Список использованных источников

1. A Deep Neural Network Model for the Task of Named Entity Recognition Anh Le, Mikhail S. Burtsev. International Journal of Machine Learning and Computing vol. 9, no. 1, pp. 8-13, 2019.
2. DeepPavlov [Электронный ресурс]. – URL: https://github.com/deepmipt/DeepPavlov (дата обращения: 15.01.2022).
3. Slovnet [Электронный ресурс]. – URL: https://github.com/natasha/slovnet (дата обращения: 15.01.2022).
4. Spacy [Электронный ресурс]. – URL: https://spacy.io (дата обращения: 15.01.2022).
5. Goal-Oriented Multi-Task BERT-Based Dialogue State Tracker Pavel Gulyaev, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, Mikhail Burtsev. AAAI - 20, 2020.
6. Маслова М. А., Дмитриев А. С., Холкин Д. О. Методы распознавание именованных сущностей в русском языке: Инженерный вестник Дона, 2021. – № 7(79) – 3–105 с.