ИТМО

Обработка текстовой медицинской информации: метод сбора и маркировки симптомов заболеваний

Подготовил: Дмитрий Погребной, 42332с

Медицинские записи

- В здравоохранении существует множество моделей прогнозирования и принятия решений
- Такие модели основываются на медицинских записях пациентов
- Точность таких моделей зависит от качества обработки медицинских записей
- Чем лучше извлекаются симптомы, тем лучше работают модели

Медицинские датасеты

- Публичные датасеты
  - RuMedNLI – 14716 records
  - RuMedPrimeData – 15249 records
- Приватные датасеты
  - Almazov National Medical Research Center – 2355 records
  - Research Institute of the Russian Academy of Sciences – 161 records
- Все датасеты были обработаны и объединены в один

Симптомы

- Сформирован набор симптомов
  - Собранные датасеты
  - Онлайн сервисы по определению заболевания по симптомам
  - Блоги и форумы
- Всего 80 различных симптомов
  - Боль в животе, озноб, удушье и другие

Маркировка симптомов

- Инструменты
  - SpaCy – NLP библиотека с поддержкой NER
  - negspaCy – библиотека для выделения отрицаний
- Метод
  - Выделяем сущности из текста с помощью ML модели
  - Определяем отрицания сущностей
  - Строим синтаксическое дерево
  - Ищем в дереве симптомы по паттернам и маркируем

Маркировка симптомов

- SpaCy плохо выделяет сущности симптомов
  - Для каждого симптома определяем набор шаблонов сущности
- Negspacy работает только с английским языком
  - Адаптируем специфичные для языка элементы
  - Используем русскую модель для определения отрицаний

Маркировка симптомов

- В сложных случаях модель ошибается – есть что улучшать
  - Специальная модель для медицинских текстов
  - Автоматическая генерация шаблонов сущностей
- Подход используется для извлечения симптомов из сообщений в медицинском боте

У пациента боль в животе, недомогание,
но нет повышенной температуры и тошнота отсутствует.

Спасибо за внимание!

ITMO more than a UNIVERSITY