Здравствуйте! Меня зовут Дмитрий Погребной. И я расскажу про архитектуру модуля языковой модели для предсказания предварительного диагноза пациента при взаимодействии с использованием диалогового агента.

Почти во всех медицинских организациях первый приём пациента начинается с того что собирается анамнез пациента и ставится предварительный диагноз на основании которого пациенту рекомендуется сдать необходимые анализы чтобы подтвердить или опровергнуть предварительный диагноз. Вот оба этих процесса требуют много ресурсов, квалифицированный персонал тратит довольно много времени на каждого пациента. И если пациентов будет довольно много, то это может перегрузить врачей, что снизит их производительность и качество лечения. Сбор анамнеза и постановка предварительного диагноза могут быть автоматизированы с помощью некоторой модели на базе машинного обучения, которая сможет взять на себя часть процессов и тем самым разгрузить врачей.

А на данный момент существует несколько готовых проприетарных решений от довольно известных компаний. Например справочные системы для поиска диагноза по заданным симптомам. Это например карта симптомов от mail.ru. Также есть примитивные диалоговый агенты на правилах и ветвлениях. Это например чат-бот сбер здоровья. А также есть проприетарное решение для определения диагноза по списку симптомов. Например сервис симптомчекер от компании webiomed.

Но во всех этих инструментах есть одно существенное ограничение. Данные решения не поддерживают работу с естественным языком. Они представляют пользователю какой-то интерфейс, но они не могут работать с естественным языком, на котором обычно привыкли разговаривать пациенты, описывать свои симптомы и болезни.

И собственно наша цель в этом проекте была разработать языковую модель и диалоговый агент, который позволит в режиме диалога собирать анамнез у пациента и предоставлять предварительный диагноз на базе сохранённых данных о болезнях и симптомах. И вот представлена диаграмма на слайде. Диалогового агент запрашивает симптомы, пациент пишет свои симптомы на естественном языке, диалоговый агент это всё обрабатывает, задаёт если необходимо уточняющие вопросы, пациент отвечает на них также на естественном языке и модель предоставляет предварительный диагноз с объяснением почему именно это диагноз наиболее вероятен.

В качестве Датасета мы используем априорные знания оболезнях и их симптомах, а также мы собрали свой датасет симптомов и болезней из форумов и других открытых источников. Этот датасет вложен в открытый доступ на платформе Kaggle и насчитывает чуть больше 5.000 записей. На диаграмме на слайде представлено распределение болезней по категориям. И например наибольшее количество болезней относится к категории дерматологии и венерологии.

На данном слайде представлена архитектура модуля. Компонента диалог-агент отвечает за взаимодействие с пациентом и составления сообщений. Компонента Preprocessor отвечает за предобоработку входящих сообщений от пользователя. Компонента Symptom Extractor отвечает за извлечение симптомов из сообщений пациента и таким образом преобразует сообщения пациента в векторное представление. Компонента Symptom Rules содержит набор правил по которым выделяются симптомы из сообщений пациента. Deisease Classifier отвечает за предсказания предварительного диагноза по векторному представлению полученному из компоненты Symptom Extractor. Symptom Explainer предоставляет объяснение, почему модель выдала именно такое предсказание и какие симптомы на это больше всего повлияли. Полученная информация возвращается в диалоговый агент и выводится пользователю как результат.

И теперь давайте я расскажу немного подробнее про сбор симптомов пациента и определение предварительного диагноза. Начнём со сбора симптомов. Собираются симптомы пациента на естественном языке, то есть пациент пишет в свободной форме что его беспокоит и мы выделяем из его сообщения симптомы, которые у этого пациента есть. Каждое сообщение мы предварительно обрабатываем, удаляем знаки пунктуации и лемматизируем слова. Для выделения симптомов мы используем библиотеку Spacy и заранее сгенерированные правила для выделения симптомов. Мы сгенерировали правила для чуть больше 5000 симптомов. Также мы стараемся обработать отрицания симптомов, то есть например когда пользователь пишет что у него нет температуры и мы должны это тоже учесть. Такие случаи мы обрабатываем с помощью библиотеки negspacy. В результате получаем векторное представление о симптомах пациента, которое потом можно передать модели машинного обучения, которая позволит на основании этого классифицировать этот набор наличия и отсутствия симптомов под некоторый предварительный диагноз.

И второй шаг после выделения симптомов это определение предварительного диагноза. Для его определения используется ML модель. Сейчас мы используем логистическую регрессию, так как для предсказания диагноза важна интерпретируемость. Нам нужно получить в резльтате не просто какой-то предварительный диагноз, а нам ещё хочется понять из-за чего такой диагноз получился, какие симптомы оказали наибольшее влияние и как-то соотнести полученные данные с реальным миром. Именно поэтому на данном этапе используется логистическая регрессия. Возможно в будущем будет что-то более сложное, например что-то на деревьях решений, но пока мы с этими экспериментируем.

Ещё важно отметить, что если модель не уверена в предсказании и не может точно определить предварительный диагноз, то у агента есть возможность уточнить наличие или отсутствие конкретных симптомов, чтобы сделать предсказание более точным. Таким образом у системы есть возможность сделать несколько попыток предсказать предварительный диагноз, каждый раз запрашивая дополнительную информацию о симптомах пациента.

Также мы уже получили первые метрики качества нашего прототипа. Точность предсказания на тестовой выборке оказалась 0.23. Это довольно низкий показатель и мы уже работаем над его улучшением. Одно из направлений это повышать количество поддерживаемых симптомов и качество их распознавания. Другое направление это экспериментировать более сложные модели для предсказания диагноза по симптомам.

В результате разработана языковая модель для определения предварительного диагноза пациента при симптомом на естественном языке. Опубликована альфа-версия python пакета с разработанным инструментом. Исходный код проекта лежит в открытом доступе на гитхабе.

В дальнейшем мы планируем замерить различные метрики качества инструмента. Поддержать работу с большим количеством симптомов. Улучшить точность распознавания симптомов. И провести эксперименты с более сложными моделями для определения симптомов.

На этом у меня всё, спасибо за внимание.