Здравствуйте, меня зовут Дмитрий Погребной. И я расскажу про обработку текстовой медицинской информации, а конкретно про метод сбора и маркировки симптомов заболеваний.

Сначала расскажу про сами медицинские записи. Сейчас сфера здравоохранения активно развивается и создается множество различных моделей прогнозирования и принятия решений. Такие модели позволяют упростить или даже частично снять тяжелую механическую работу с врачей. Например такие модели могут советовать план лечения или рекомендовать какие дополнительные анализы нужно сдать пациенту. Помимо самой рекомендации такие модели также могут предоставлять объяснение своих выводов. Это позволяет понять почему модель сделала то или иное решение.

Большинство медицинских моделей основываются на электронных медицинских записях пациентов. Такие записи обычно представляют собой простой не структурированный текст, который трудно поддается обработке. Точность и корректность работы медицинских моделей во многом зависит от того насколько хорошо обрабатываются тексты. В данном же случае насколько хорошо из текста извлекаются симптомы пациента.

Для того чтобы извлекать симптомы пациентов из медицинских текстов сначала необходимо было собрать данные. Медицинских датасетов на русском языке очень ограниченное количество. Поэтому было собраны все датасеты что есть на данный момент. А конкретно это два публичных датасета и два приватных датасета. Один из приватных датасетов был предоставлен исследовательским центром имени алмазова, а второй датасет предоставлен национальным институтом российской академии наук. Полученные датасеты были обработаны. Все датасеты были лемматизированы и очищены от стоп слов. В результате все датасеты были объединены в один финальный.

Из полученного датасета, а также из онлайн сервисов по определению заболевания по симптомам и блогов и форумов был сформирован набор основных симптомов, которые необходимо будет выделять из текстов в дальнейшем. Всего получилось 80 различных симптомов. Например боль в животе, озноб, удушье и другие.

После определения списка симптомов можно начинать их маркировку для маркировки симптомов используется инструменты SpaCy и NegSpaCy. SpaCy это библиотека для работы с текстом на языке Python, которая помимо всего прочего позволяет выделять сущности из текстов, так называемое Name Entity Recognition, а также может строить синтаксические деревья предложений. NegSpaCy - это библиотека для определения отрицания выделенных сущностей в тексте.

Общий метод маркировки симптомов следующий. Сначала мы выделяем сущности из текста с помощью ml-модели и библиотеки SpaCy. Далее определяем есть ли у выделенных сущностей отрицание. Затем строим синтаксическое дерево предложения и в этом дереве по паттернам ищем симптомы и маркируемых их.

Однако к сожалению всё не так просто с этим подходом. Модель SpaCy плохо выделяет сущности симптомов и поэтому для каждого симптома необходимо определять кастомный набор шаблонов сущности. Кроме того библиотека negspaCy работает только с английским языком поэтому необходимо было адаптировать специфический для языка элементы и переписать часть для работы с русским языком, а также использовать русскую модель для определения отрицаний.

Несмотря на в целом достойное качество определение отрицания и выделения симптомов. Данный метод иногда дает некорректный результат и здесь есть что улучшить. Например можно натренировать специальную модель для медицинских текстов, что повысит точность выделения сущностей симптомов, а также можно написать автоматическую генерацию шаблонов для симптомов, что также повысит точность выделения симптомов. Описанный подход используется для извлечения симптомов и сообщений в медицинском боте и вот внизу представлен пример выделения симптомов из предложения. Зелёным обозначены утверждение симптомов, а красным отрицание симптомов. На этом у меня всё, спасибо за внимание!