ИЗВЛЕЧЕНИЯ ЗНАЧИМЫХ ПРИЗНАКОВ ИЗ ЭЛЕКТРОННЫХ МЕДИЦИНСКИХ ЗАПИСЕЙ НА ОСНОВЕ КОНТЕКСТНО-СВОБОДНОЙ ГРАММАТИКИ

С.В. Аксёнов, к.т.н., доц. ОИТ ИШИТР,
Р.Р. Котюбеев, студент, гр. 8ПМ9И
Томский политехнический университет
E-mail: rrk8@tpu.ru

Введение

Врачи работают с высокой степенью умственной нагрузки. Поэтому было бы естественно как-то эту нагрузку уменьшить. В данной работе мы предлагаем способ извлечения значимых признаков из электронных медицинских записей на основе контекстно-свободной грамматики. Так, сплошной текст с первичным осмотром мы преобразуем в структурированный вид, который позволит врачу быстро определить тот или иной признак: имелся ли у пациента озноб, какая у него была температура во время поступления, какой диагноз, зоны поражения болезнью и т.д.

Парсер Yargy

Для извлечения признаков применяется парсер Yargy, написанный на языке Python, для извлечения структурированных данных из текста на естественном языке. В основе лежит использование контекстно-свободной грамматики. Причем, парсер позволяет писать свои грамматики.

В Yargy есть банк готовых грамматик, которые позволяют найти в тексте даты, имена, организации, адреса, номера телефонов и. т.д. Для медицинских текстов этих грамматик недостаточно, поэтому мы составили свои.

Правила в Yargy могут состоять из других правили предикатов. Предикат – это функция, которая принимает на вход токен и возвращает True, если соответствующий факт был найден, и False, если не был найден. Правила и предикаты могут логически комбинироваться при помощи логических операторов and_, or_и not_.

Извлечение признаков

Целевые признаки были определены и подготовлены врачами СибГМУ. Всего таких признаков насчитывается 47. К ним относятся: возраст, провоцирующие факторы, аллергия на лекарственные препараты, сопутствующие заболевания, зоны поражения болезни, основной и сопутствующий диагноз и т.д.

Текст первичного осмотра врачом состоит из отдельных блоков. Всего таких блоков 10:

1. Общая информация о пациенте (пол, возраст, время поступления).
2. Дата и время осмотра.
3. Жалобы.
4. Анамнез болезни.
5. Анамнез жизни.
6. Эпидемиологический анамнез.
7. Анамнез ВТЭ (венозная тромбоэмболия)
8. Объективный статус.
9. Локальный статус.
10. Диагноз

Каждый из блоков содержит свой перечень признаков, который нужно извлечь. Мы также написали программу для деления на соответствующий блоки так, что нахождение признаков осуществляется в асинхронном режиме для большей скорости выполнения.

Для определения *других заболеваний* в анамнезе был составлен отдельный словарь со списком болезней. Заболевания, которые перенес пациент, указываются в таких блоках медицинской карты, как анамнез болезни, анамнез жизни и эпиданамнез. Yargy для того, чтобы не перепутать заболевания с диагнозом также предварительно нормализует слова в них, а затем находит заболевания.

Поскольку в первичном осмотре не указывается отсутствие каких-то симптомов, для их определения достаточно было передать список значимых симптом в парсер без создания сложных правил.

Yargy имеет синтаксический анализатор Pymorphy2. С помощью него можно определять часть речи слова, поэтому были созданы правила, в которых после определенной фразы ожидается та или

иная часть речи. Такие правила позволили определить препараты, вызывающие у пациента аллергическую реакцию. Таким образом, Парсер Yargy искал фразу “Аллергическая реакция на” и после нее определял набор существительных и прилагательных разделенных запятой как препараты. Соответствующее правило реализовано следующим образом:

```
ALLERG_RULE= or_(
    rule(
        normalized('Аллергическая'),
        normalized('реакция'),
        normalized('на'),
        gram('NOUN').optional().repeatable(),
        gram('ADJF').optional().repeatable()),
    rule(
        normalized('Аллергическая'),
        normalized('реакция'),
        normalized('на'),
        gram('ADJF').optional().repeateble(),
        gram('NOUN').optional().repeatable())
)
```

Предикат *normalized* приводит к начальной форме слова. Метод *optional* указывает на то, что предикат может появиться, а может нет. Метод *repeatable* указывает, что предикат в тексте может повторяться.

Тестирование парсера с созданными правилами

При тестировании использовалось 36 записей первичного осмотра врачом. В каждом из них определены 49 признаков. Таблица 1 показывает точность определения признаков: 44 были определены с высокой точностью, признак «кем направлен» на 75 %, 4 признака определены с точностью 70 %.

Заключение

В результате были разработаны алгоритм для извлечения значимых признаков из электронных медицинских записей. Данный алгоритм показывает высокие показатели на 44 признаков из 49. В дальнейшем планируется увеличить точность на остальных 5 признаках.

Список использованных источников

1. Амосов О.С.Zdrav Expert — электронные медицинские карты (ЭМК) [Электронный ресурс]. - URL: http://zdrav.expert/index.php/Статья:Электронные_медицин-ские_карты_(ЭМК)#2019 (дата обращения:02.03.2021).
2. Документация Yargy парсер [Электронный ресурс]. - URL: ttps://tech.yandex.ru/tomita/ (дата обращения:02.03.2021).