УДК 004.89

ИЗВЛЕЧЕНИЕ СИМПТОМОВ ИЗ ЭПИКРИЗОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Галай О.О. (ИТМО)
Научный руководитель – кандидат технических наук Русак А.В.
(ИТМО)

Введение. Рынок HealthNet входит в матрицу Национальной Технологической Инициативы (НТИ). Одним из ее сегментов является развитие информационных технологий в медицине, среди которых выделяют устройства и сервисы для мониторинга и коррекции состояния человека [1]. В рамках мониторинга можно выделить интеллектуальную обработку электронной персональной медицинской записи (ЭПМЗ). Метод извлечения симптомов из эпикризов решает одну из задач интеллектуальной обработки медицинских данных.

Основное содержание. Решение задачи извлечения симптомов из медицинских записей с использованием методов машинного обучения можно разделить на следующие этапы: предварительная обработка данных, определение признакового пространства, обучение модели и постобработка полученных данных для выделения координат последовательностей в тексте. В качестве входных данных использовался датасет, состоящий из xml-файлов эпикризов и json-файлов эталонных разметок, в количестве 145 тренировочных и 99 тестовых.

Предварительная обработка данных подготавливает данные для последующего обучения модели. Задача извлечения симптомов схожа с задачей извлечения именованных сущностей, которая во многих исследованиях преобразуется в задачу классификации с использованием IOB-разметки [2][3]. Она представляет собой присвоение каждому токену текста одного из трех классов. Классы представлены следующим образом: I (inside) – внутренняя часть сущности, в данной задаче симптома, O (outside) – не является сущностью или B (beginning) – начало сущности. Данные классы были присвоены токенам, извлеченным из текстов секций эпикриза, на основании данных из эталонной разметки.

Следующим шагом было выделено признаковое пространство. Оно включает в себя следующие признаки: лемма токена, лемма до последних трех букв, флаг, состоит ли токен только из символов верхнего регистра, флаг, начинается ли токен на символ верхнего регистра, код медицинской сущности токена, часть речи токена. Также в случае, если токен не является начальным в секции файла, то такие же признаки собираются для предыдущего токена, и аналогично, если токен не является конечным в секции файла, - для последующего. Данный подход позволяет обрабатывать токен с использованием данных о соседних токенах [4].

Предсказание классов для каждого токена производилось с использованием классификатора CRF (conditional random field или условное случайное поле) [2][4]. Данный классификатор является классическим решением задачи классификации последовательностей. Его преимущество перед дискретными классификаторами заключается в том, что он способен учитывать контекст. Также CRF позволяет передавать признаки различной длины, что является несомненным плюсом в обработке естественного языка. В результате для каждого токена тестового датасета был получен класс IOB-разметки.

Для перехода от классовой разметки к координатам симптомов в тексте эпикриза был разработан алгоритм, производящий поиск с использованием регулярных выражений, в результате которого по переданным индексам последовательности токенов возвращается координата начала и конца симптома.

Выводы. В результате подбора признакового пространства, поиска параметров для алгоритма CRF, а также извлечения координат последовательностей токенов был получен

алгоритм, получающий на вход документ xml-формата и возвращающий json-файл, содержащий список найденных в тексте документа симптомов. Для оценки точности алгоритма была использована метрика оценки точности последовательностей seqeval и получена F1-мера 0.722, для каждого класса по отдельности F1-мера достигла следующих значений: 0.825 – для класса «B», 0.853 – для класса «I» и 0.950 – для класса «O».

Список использованных источников:

1. Протокол №1 заседания Межведомственной рабочей группы по разработке и реализации Национальной технологической инициативы при Правительственной комиссии по модернизации экономики и инновационному развитию России от 21 января 2021 г.
2. K. Jia, W. Weiji, C. Xiaojun, G. Jianping, G. Yan, J. Shuai. Medical entity recognition and knowledge map relationship analysis of Chinese EMRs based on improved BiLSTM-CRF // Computers and Electrical Engineering. — 2023. — № 108.
3. A. Nasser, A. Saad. The impact of using different annotation schemes in named entity recognition // Egyptian Informatics Journal. — 2021. — № 22. — С. 295-302.
4. S. Richa, M. Sudha, A. Basant. Named entity recognition using neural language model and CRF for Hindi language // Computer Speech & Language. — 2022. — № 10.