УДК 004.891.3

**Исследование и разработка модели и алгоритма формализованного представления историй болезни пациентов на основе неструктурированных данных**

*И.Е. Постоев, А.М. Андрианов*

*Национальный исследовательский университет «МИЭТ»*

e-mail: [postoev.ie@gmail.com](mailto:alyonakhustochka@gmail.com)

Эпилепсия — это тяжёлое заболевание головного мозга, характеризующееся периодическими нарушениями деятельности мозга, возникающими вследствие неконтролируемой активности нейронов головного мозга. Лечение эпилепсии продолжительностью 5-7 лет и сопровождается подбором методов в зависимости от анализа ЭЭГ и описаний приступов, которые составляются самими пациентами или их родственниками в неструктурированном виде. Частота приступов может составлять до нескольких эпизодов в день, в то время как исследование с помощью ЭЭГ из-за высокой стоимости производится сравнительно редко: раз в 3-4 месяца. Таким образом, описания приступов являются важным источником данных для врача при назначении и корректировки плана лечения. На сегодняшний день врачи должны вручную извлекать из текстов описаний полезные данные для внесения в историю болезни и принятия решений о лечении, что является трудоёмкой операцией.

Время проведения анализа описаний эпизодов было бы существенно уменьшено в случае, если данные описания были бы представлены в формализованном виде, согласно правилам оформления истории болезни и содержащими лишь актуальные для специалиста данные. В них бы фигурировали лишь те полезные данные (далее - сущности), которые являются значимыми согласно заранее определенным критериям. Данную задачу можно переформулировать как задачу извлечения полезных сущностей из неструктурированных данных эпилепсии (НДЭ), выраженных естественным языком. Она является представителем класса задач распознавания имён сущностей (NER), являющегося подклассом задач по обработке естественного языка(NLP).

Существует множество направлений решения задачи NER. Среди них - системы, основанных на правилах (rule-based, например - регулярные выражения), которые ограничены сложностью описаний правил. Методы машинного обучения (ML-методы - например, условные случайные поля - CRF) более гибкие - правила извлечения сущностей подбираются в ходе обучения системы. Выбор того или иного метода зависит от задачи, для решения которой система разрабатывается и анализируемыми входными данными. Так, в исследовании эффективности данных методов[1] было определено, что персональные данные пациентов обработанные системами, использующими оба основных типа методов, были извлечены с высокой точностью (рис.1). Эффективность алгоритма в данном случае определяется специальными мерами, используемыми в задачах NER: P(Precision) – точность, R(Recall) – полнота, F(F-score) – F-мера, A(Accuracy) – доля правильных ответов алгоритма. Испытуемыми системами являются: Rule – rule-based система, ML1 – ML-система, обученная на исследовательском корпусе данных MedNLP и ML2 - аналогичная система, обученная на медицинских статьях из открытых интернет-источников.

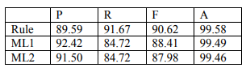


Рис.1 Точность извлечения персональных данных

При оценке данных о состоянии пациента, которые, как правило, менее структурированы, методы ML имеют преимущество перед rule-based (рис. 2).

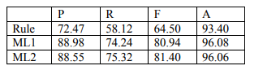


Рис.2 Точность извлечения данных о состоянии

Исходя из этого, для обработки НДЭ более подходящим будет использование ML-методов или их комбинация с rule-based методами.

При имеющейся задаче извлечении сущностей следует учитывать действие закона Ципфа, согласно которому частота встречаемости каждого слова достаточно длинного текста примерно обратно пропорциональна порядковому номеру (рангу) этого слова в списке слов, упорядоченному по частоте встречаемости. Пример данной зависимости показан на рис. 3.

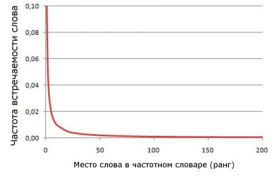


Рис 3. График зависимости частоты встречаемости слова от его ранга

Негативное влияние закона на итоговую эффективность работы нейросетей заключается в том, что при их обучении на размеченных вручную корпусах данных, имеющих ограниченные размеры, могут быть упущены многие слова, имеющие высокий ранг и, соответственно, неправильно распознанные в ходе работы нейросети. Для уменьшения влияния данного эффекта были использованы алгоритмы, проводящие обучение на неразмеченных корпусах данных, имеющих большие размеры. Предполагается использование данных методов в алгоритме, используемом в разрабатываемом решении.

После выделения необходимых критериев поиска, определяется формат исходных и результирующих данных, который будет наиболее эффективен для анализа специалистом и дальнейшей возможной обработки полезных данных в других решениях. После, исходя из результатов экспериментального исследования, будет определена функциональная схема алгоритма и проведён поиск и составление обучающих корпусов данных для нейронной сети. Обучающие данные максимально приближены к данным, которые будут обрабатываться в ходе решения реальных задач и составлены определенным в зависимости от выбранного алгоритма обучения, образом. В завершение обучения нейронной сети, разрабатывается программное обеспечение для оценки эффективности работы алгоритма и проведение тестовых испытаний с постепенным внедрением в процесс работы согласно требованиям специалистов. Результатом работы обученной нейронной сети является формализованное описание по правилам заполнения истории болезни, удобное для анализа и принятия врачом решения об изменениях в лечении.

Задача извлечения полезных сущностей из неструктурированного текста медицинской тематики не является новой – был реализован ряд решений по анализу текстов медицинских карт пациентов, выделению в них медицинских сущностей. При этом, на данный момент в медицинских учреждениях РФ в ходе регулярной работы не используются решения, имеющие схожий функционал. Известными продуктами, реализующими его являются системы «KnowledgeMap», «MedLEE», «MetaMap». Данные решения представляют реализацию систем, основанных на правилах, что подразумевает ограничения в точности распознавания. Примерами же систем, в которых используются методы машинного обучения, и, соответственно, имеющие большую точность, являются англоязычный «CLAMP» и способный работать с текстом на русском языке «Droice Labs». Однако данные решения работают с текстами, заполненными специалистами - медиками, в то время как описания эпизодов эпилептических приступов осуществляют сами пациенты или их родственники.

Таким образом, разработка решения для обработки неструктурированных описаний эпизодов эпилептических приступов является востребованным как в новом исследовании применения алгоритмов обучения нейросетей к задаче, имеющей определённую специфику, так и в качестве инструмента для повышения эффективности работы специалистов области эпилепсии.

1. Osamu Imaichi, Toshihiko Yanase, and Yoshiki Niwa A Comparison of Rule-Based and Machine Learning Methods for Medical Information Extraction Hitachi, Ltd., Central Research Laboratory.