Минобрнауки России

Федеральное государственное автономное образовательное учреждение высшего образования   
«Национальный исследовательский университет   
«Московский институт электронной техники»

Институт cистемной и программной инженерии и информационных технологий

(СПИНТех)

Постоев Игорь Евгеньевич

Магистерская диссертация   
по направлению 09.04.04 «Программная инженерия»

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

Студент Постоев И.Е.

Руководитель,   
к.т.н., доцент Андрианов А.М.

Москва 2019

# Список сокращений

NLP – Native Language Processing

NER – Named Entity Recognition

НДЭ – неструктурированные данные эпилепсии

# Введение

### Актуальность исследования

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 неструктурированного текста, написанного на естественном языке. Она является типичным представителем класса задач распознавания имен сущностей (NER), являющееся подклассом более общего свода задач по обработке естественного языка(NLP).

### Объект исследования

Объектом исследования диссертационной работы являются неструктурированные тексты, описывающие протекание эпилепсии у пациентов.

### Предмет исследования

Предметом исследования диссертационной работы является алгоритм извлечения полезных данных из НДЭ.

### Цели и задачи исследования

Целью данной работы является получение создание алгоритма для эффективного извлечения данных из НДЭ.

В соответствии с целью и предметом исследования в диссертации неоходимо решить следующие задачи:

* определить основные проблемы и ключевые критерии оценки эффективности решения.
* исследовать существующие методы и алгоритмы NER

с целью выявления их характеристик эффективности, преимуществ и недостатков;

* произвести формализованную постановку задачи извлечения данных из НДЭ;
  + разработать функциональную схему алгоритма формализованного представления историй болезни на основе НДЭ;
* провести экспериментальное исследование алгоритма;
* создание тестовой модели на прежде сформированном корпусе данных для обучения;
  + разработать программное обеспечение для оценки эффективности алгоритма;
  + провести оценку результатов исследования, сравнительный анализ результатов разработки и экспериментов с существующими решениями.