**Отчёт обучающегося по практике**

Обучающийся: Постоев Игорь Евгеньевич Группа: ПИН-12М

Направление: 09.04.04 «Программная инженерия»

Образовательная программа: Программная инженерия искусственного интеллекта

Вид практики: учебная

Осенний семестр 2019/2020 учебного года

Место практики: АО «Системы управления»

`

СОДЕРЖАНИЕ

[Введение 2](#_Toc28067303)

[Применение извлечения именованных сущностей 3](#_Toc28067304)

[Выбор языка программирования 5](#_Toc28067305)

[Структура разрабатываемого ПО 5](#_Toc28067306)

[Результаты работы программы 8](#_Toc28067307)

[Заключение 11](#_Toc28067308)

[Список использованных источников 12](#_Toc28067309)

**Введение**

Специалисты области эпилепсии на данный момент не имеют инструмента формализованного представления данных о протеканиях эпилептических эпизодов, что существенно повысило бы производительность их работы и эффективность принятыми ими решений. В обработанных, формализованных данных должны фигурировать 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), являющегося подклассом более общего свода задач по обработке естественного языка (NLP).

Существует множество различных задач, в которых применяется системы, способные извлекать и структурировать полезные данные согласно заданным критериям выборки, жанру текста извлекаемой информации (научные статьи, неформальная, повседневная речь, новостные сводки, художественный или медицинский текст), типам извлекаемых сущностей (например, локации, персоны, организации, параметры некоторых конкретных объектов). Для подобных задач, имеющих набор характерных признаков, существует ряд подходов к её решению и возможность создания системы для эффективного извлечения именованных сущностей. Такие подходы включают скрытые марковские модели (Hidden Markov Models, HMM), решающие деревья(Decision Tree, DT), условные случайные поля(Conditional Random Field, CRF), методы максимальной энтропии(Maximum Entropy, MaxEnt), опорных векторов(Support Vector Machine, SVM) и другие, каждый из которых имеет свои достоинства и

недостатки [1]. В данной работе проводится сравнительный анализ методов решения задачи NER, определение наиболее эффективных для поставленной задачи извлечения сущностей из неструктурированных данных эпилепсии.

**Применение извлечения именованных сущностей**

Извлечение именованных сущностей находит применение во многих

областях, связанных с обработкой текстов на естественных языках и извлечением информации. Рассмотрим некоторые из них:

* Информационный поиск

Классическая задача информационного поиска подразумевала поиск

документа, удовлетворяющего критериям запроса, среди коллекции документов. Сейчас данная задача включает такие подзадачи, как классификацию, фильтрацию, кластеризацию документов, аннотирование и реферирование документов, создание и совершенствование языков запросов. Извлечение именованных сущностей позволяет проводить индексацию и поиск документов более эффективно.

* Вопросно-ответные системы

Вопросно-ответная система принимает вопросы и дает на них ответы на естественном языке. Большая часть вопросов подразумевает в качестве ответа то, что мы относим к именованным сущностям различных типов. Что делает применение извлечения именованных сущностей при построении подобных систем полезным с точки зрения увеличения их производительности и качества работы.

* Автоматизированный сбор новостей

Новости посвящены некоторым событиям. События, описываемые в

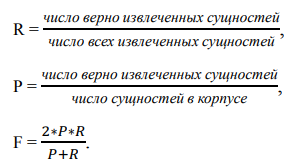
новостях, можно ёмко охарактеризовать с помощью различных именованных сущностей – место действия время действия участники события .

* Анализ биологических и медицинских текстов

Выделение специализированных типов именованных сущностей, таких как названия белков, генов, областей действия генов, типов клеток и организмов в области биологии и названия заболеваний, лекарств, действующих веществ в области медицины позволяет проводить более эффективный поиск информации среди огромного количества документов, относящихся к данным областям.

### Методы оценки систем извлечения именованных сущностей

Обычно оценка системы извлечения именованных сущностей проводится на размеченном вручную корпусе. Именованная сущность определяется своими границами (входящими в нее словами) и типом. Применяются различные способы измерения и оценки. На конференции CoNLL был предложен следующий способ оценки. Если границы и тип сущности совпадают с разметкой эксперта, содержащейся в корпусе, то сущность считается извлеченной верно, иначе имеет место ошибка системы. Такой способ оценки называется оценкой методом точного соответствия. Этот способ получил широкое распространение, однако имеет и свои недостатки: он не учитывает возможные ошибки, совершенные экспертом при разметке. Например, можно выделить дополнительные типы ошибок: верно определен тип, но ошибочна граница, верно выделена граница, но ошибочен тип, и неверно определены и тип сущности, и граница. На конференциях MUC системы оценивались параллельно по двум направлениям: правильности выделения границ именованных сущностей и правильности определения типов. К показателям качества работы системы относят полноту(Recall, R), точность(Precision, P) и F-меру(F), которые рассчитываются следующим образом:



Извлечение именованных сущностей являлось задачей на многих конференциях, в частности, CoNLL и MUC. В связи с этим существуют некоторые готовые корпуса для систем извлечения именованных сущностей.

### Способы кодирования слова в рамках именованной сущности

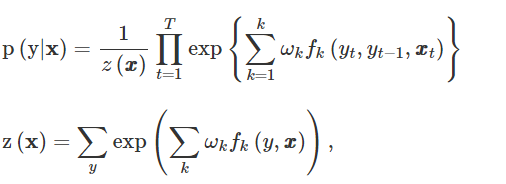
При кодировании способом BIO различают начальные (Beginning), внутренние (Inner) и внешние (Outer) части именованной сущности. Таким образом, слово, являющееся первым в последовательности слов, обозначающих именованную сущность, получит дополнительную метку B, все последующие слова последовательности, относящиеся к этой же именованной сущности – метку I, слово последовательности, которое не относится к именованной сущности – метку O. При кодировании способом BILOU различают начальные (Beginning), внутренние (Inner), завершающие(Last) и внешние (Outer) части именованной сущности, состоящей из нескольких слов, а также именованную сущность, представляемую одним словом(Unit).

Метод машинного обучения с учителем заключается в необходимости построить и обучить модель, которая будет решать задачу классификации. Объект представляет собой набор признаков. На этапе обучения модели классификатор принимает пары «набор признаков – класс», которые описывают принадлежность объекта, характеризуемого данным набором признаков, к данному классу. Далее классификатор в соответствии с используемыми алгоритмами и эвристиками устанавливает зависимости между значениями тех или иных признаков (и их комбинациями) и принадлежностью объекта к тому или иному классу. На этапе тестирования и функционирования на вход подается только набор признаков, а классификатор, руководствуясь выявленными на этапе обучения зависимостями, относит объект к некоторому классу. На вход системе извлечения именованных сущностей подается текст как упорядоченный набор слов. Аннотировать его метками – значит отнести каждое слово к некоторому классу. Таким образом, объектом, набор признаков которого подается на вход классификатору, является слово, а классом – метка. Значит, при решении задачи извлечения именованных сущностей 4 типов(PER, ORG, LOC, MISC) и использовании кодирования BILOU имеем 17 классов (с учетом метки О, которая, как говорилось выше, обозначает, что слово не относится к именованной сущности). Значения признаков, как правило, можно представить числами. Так, например, булевское значение признака «начинается с заглавной буквы» можно представить 0 или 1.

### Методы решения задачи NER

Метод условных случайных полей (CRF) - это модель моделирования статистической последовательности. CRF представляют собой класс методов статистического моделирования, часто применяемых в распознавании образов и машинном обучении. В то время как обычный классификатор предсказывает метку для одной выборки без учета «соседних» выборок, CRFспособен учитывать контекст.

В этой модели X = {x1, x2, x3,… xT} являются входными данными, в которых компоненты соединены последовательно, а Y = {y1, y2, y3,… yT} являются метками для каждого компонента входных данных. данные. Другими словами, когда производится новый x, значение y прогнозируется с использованием следующей модели:



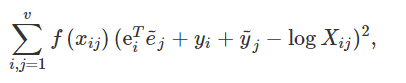
где z (x) стандартизирует значение вероятности, а fk является функцией признака, которая является характеристической функцией признака k. Эта функция возвращает 1, когда заданный вход yt, yt − 1, xxt включает в себя функцию k, и возвращает 0 в противном случае. ωk - вес элемента.

Встраивание слов(Word embedding)

Вложение слова также называется представлением слова или распределенным представлением. Он изучает векторное представление для каждого слова, встречающегося в корпусе. Предыдущие исследования представляли собой слово как поиск по словарю. Данное представление использует вектор размером с словарь и принимает 1, когда слово появляется в документе, и 0, когда его нет. Вложение слов уменьшает размеры и разреженность исходного вектора и заполняет вектор действительными числами.

Глобальные векторы(GloVe)

GloVe - это неконтролируемый алгоритм обучения для получения векторных представлений для слов. Обучение выполняется на агрегированной глобальной статистике совпадений слово-слово из корпуса, и полученные представления демонстрируют интересные линейные подструктуры векторного пространства слов. GloVe учитывает как глобальный контекст, так и локальный контекст .

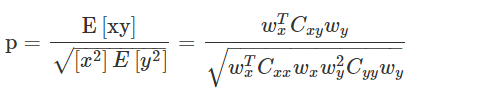


где X - это слово, встречающееся в матрице, Xij - частота совпадения слов i и j, а Xi = ∑vkXik - общее количество вхождений слова i в корпус. Вероятность появления слова j в контексте слова i равна Xij = P (j | i) = Xij / xi. e - вложение слов, а e - вложение слов в отдельном контексте.

Канонический корреляционный анализ (CCA)

Канонический корреляционный анализ является статистическим методом для исследования взаимосвязи между двумя наборами переменных, и он может одновременно исследовать корреляцию переменных, принадлежащих к различным наборам. CCA находит корреляции между двумя наборами переменных (X, Y), а также находит параметры, которые максимизируют коэффициенты корреляции. CCA может быть рассчитана непосредственно из набора данных, а также может быть рассчитана после преобразования наборов данных в ковариационные матрицы.

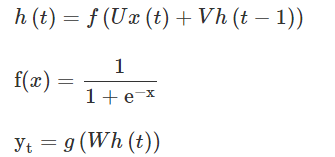
Предполагая, что у нас есть две переменные x∈Cd1, y∈Cd2, CCA, можно определить как проблему, чтобы максимизировать корреляцию между двумя переменными на векторах X и Y. С парой векторов x = w ^ Tx, y = w ^ Ty мы можем использовать следующее выражение корреляции:



Рекуррентная нейронная сеть

В машинном обучении и когнитивной науке искусственные нейронные сети (ANN) представляют собой семейство моделей, основанных на биологических нейронных сетях, которые используются для оценки или аппроксимации функций, которые могут зависеть от большого количества входных данных и обычно неизвестны. ANN хорошо работают в нелинейных функциях и распознавании образов. Многие исследователи, работающие в области интеллектуального анализа данных, искусственного интеллекта и биоинформатики, интересовались ANN для его разнообразных применений.

RNN имеет два типа: сеть типа и сеть типа Джордана . Сеть типа Эльмана добавляет уровень контекста к обычному RNN и передает выходные данные скрытого уровня на вход уровня контекста. Эта сеть возвращает выходное значение скрытому слою, а не входному слою. Скрытый уровень этой сети играет ту же роль, что и входной уровень обычного RNN. На рисунке 4 показана базовая структура РНН типа Эльмана. Выходные данные скрытого слоя, сигмовидная функция каждого узла и выходное значение этой сети представлены ниже:



**Заключение**

За время прохождения учебной практики были исследованы основные методы решения задачи извлечения сущностей из неструктурированного текста.

Во время прохождения учебной практики были освоены следующие компетенции и подкомпетенции:

УК-1. Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий.

ПК-2. Владение навыками создания программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов.

ОПК-3. Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями.

**Список использованных источников**

1. Clinical Named Entity Recognition Using Deep Learning Model [Электронный ресурс]. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977567/
2. Introduction to Named Entity Recognition [Электронный ресурс]. – URL: <https://medium.com/explore-artificial-intelligence/introduction-to-named-entity-recognition-eda8c97c2db1>
3. Comparative Analysis between Notations to Classify Named Entities using Conditional Random Fields. Daniela Oliviera F. do Amaral, Maiki Buffet, Renata Vieira. 1Faculdade de Informatica – Pontif ´ ´ıcia Universidade Catolica do Rio Grande do Sul (PUCRS) ´ Caixa Postal 1429 – 90619-900 – Porto Alegre – RS – Brazil
4. Comparison of named entity recognition methodologies in biomedical documents [Электронный ресурс]. – URL: <https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-018-0573-6>

Руководитель практики от МИЭТ / Федоров А.Р./

Руководитель практики   
от организации / Муравин В.Б./

Обучающийся / Постоев И.Е./