Сводка по написанию дипломной работы

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

Медицинские анамнезы эпилепсии содержат необходимые данные для формирования специалистом способа лечения. Значащие, полезные данные обычно представлены в неструктурированном виде, при наличии незначащей информации, и требуют формализации. Структурированное представление данных может помочь специалисту быстрее ознакомиться с нужными параметрами, что может увеличить эффективность работы. Более того, данные, представленные в формальном виде, могут иметь различное применение, например, для выявления причинно-следственной связи между различными событиями, условиями, значащими при эпилепсии. Основная задача при такой формализации данных – выделить в «сыром» тексте анамнеза интересные специалисту данные. Данная задача является задачей обработки естественного языка и называется извлечением именованных сущностей(англ. – NER). Под извлечением сущности понимается сопоставление слову соответствующего ему класса(тега) на основании его значения и контекста. Критерием качества решения задачи является точность соотнесения слов классам. Задача NER для медицинских текстов на русском языке исследована недостаточно, а доказавших эффективность готовых решений нет. В области эпилепсии подобных исследований не проводилось, и, соответственно, поставленная задача является актуальной.

Существует несколько подходов к решению такой задачи. Самый простой, но наименее эффективный – использование заранее составленных словарей, что не позволяет учесть контекст предложения. Возможно составление регулярных выражений, что также вносит существенные ограничения на контекст слова. В таком случае прибегают к статистическим методам.

Для применения статистических методов, слова текста требуется представить в скалярном или векторном виде. Однако, так как слово может иметь много форм за счет аффиксов, окончаний, сначала требуется его нормализация. Обычно применяют лемматизацию – получение леммы слова. Для лемматизации и NER также проводится POS-тегирование – определение части речи слова. Далее слова эффективно представлять в виде эмбеддингов – векторов, характеризующих семантическую близость слов. Для этого используют алгоритм word2vec и его модификации.

В статистических генеративных методах определяется наиболее вероятная последовательность тегов. Так, в методе N-грам, который основан на идее теоремы Байеса, на вероятность соотнесения слова тегу влияют N результатов для предыдущих слов. Генеративные методы включают скрытые марковские цепи(HMM), где моделью является автомат и последовательность тегов строится при переходе между состояниями этого автомата. Нейросетевая модель RNN и ее модификации – LSTM, BiLSTM – также принадлежат этой группе алгоритмов. Идея RNN – использование одних и тех же весов при обучении на последовательности слов. Обучается методом обратного распространения по времени. В качестве модификации применяется механизм внимания, чтобы на соотнесение слова мог иметь наибольшее влияние конкретный участок анализируемого текста, а не усредненное значение для предыдущих слов (скрытое состояние). Недостатками этой группы методов являются сложность применения параллельных вычислений и проблема зависимости расположенных на отдалении слов.

Основное назначение дискриминативных моделей – классификация. Одним из наиболее используемых алгоритмов данной группы является метод условных полей(CRF). CRF показывает результат для задачи NER по мере F1 лучше чем HMM, но обладает высокой сложностью обучения, и может уступать по точности многим методам на больших датасетах.

Архитектура Transformer на данный момент является наиболее актуальной – используется в популярных переводчиках и средствах анализа текста. Трансформеры избавлены от недостатка генеративных алгоритмов – предложения анализируются целиком, а не по отдельному слову. Трансформеры показывают эффективность при обучении их на больших корпусах данных. На основе такой архитектуры был создана модель BERT. Существует возможность «дообучить» BERT для решения специфичной задачи машинного обучения.

Таким образом, есть идея модифицировать BERT с помощью алгоритмов NER, комбинируя различные методы применимо к классификации слов в тексте эпилепсии. То есть, например, дообучить BERT с помощью LSTM и CRF применяя собственный датасет(не важно, сформирован ли он из «реальных» данных или синтетически). Также, для объективной оценки следует применить комбинации методов NER и без использования BERT. Проверить точность модели на тестовых данных, и сделать вывод о применимости той или мной комбинации алгоритмов. В итоге, определить такой метод, при котором достигалась бы лучшая точность с учетом ограниченности данных для обучения и специфичности данных эпилепсии.