Сводка ко второй главе дипломной работы

В целях формализации задачи, необходимо определиться с методом извлечения именованных сущностей. Чтобы определить, с помощью каких методов потенциально можно было бы достичь большей точности в распознавании, следует обратиться к результатам испытаний в решении задачи NER с целью выявления алгоритмов, обладающих наибольшей точностью.

Наиболее популярным критерием точности классификации слов является f-мера. Данная метрика является гармоническим средним точности и полноты:

, где

precision - мера, характеризующая правильность соотнесения слов тому или иному классу, а recall характеризует, насколько полным было данное соотнесение.

Как было указано в главе 1, существуют различные методы и подходы к задаче NER. При выборе алгоритма решения руководствуются спецификой задачи и экспериментальными данными тестирования алгоритмов на схожих корпусах данных и при решении схожих задач(в данном случае - NER).

Специфика задачи извлечения данных из НДЭ следующая:

* наличие профессионализмов, терминов, которые могут встречаться редко;
* требования к извлекаемым сущностям могут быть композитные, например, определить частоту приступов в настоящий момент, или после приема определенного препарата, или при наличии/отсутствии других условий, что говорит о возможном большом количестве извлекаемых сущностей;
* вместе с наличием специфичных для эпилепсии терминов, в анамнезах большое количество слов из общей медицины или общеупотребимых фраз, которые тоже могут нести важный смысл в контексте;
* размер исходного датасета для обучения относительно небольшой;

F-score =

Задача NER условно может быть разделена на 3 этапа:

- представление входных данных;

- составление контекста (предложения, документа, другой области текста) на основе входных данных – кодирование;

- классификация слов на основании контекста – декодирование;

Представление данных

Вектор входных данных может включать информацию символьного уровня, уровня слова и гибридную информацию(возможные комбинации двух предыдущих представлений, уровня части речи слова, BIO-тега, глобального(word2Vec, GloVe) и контекстуального эмбеддинга слова(BERT), и др.). Согласно <https://arxiv.org/pdf/1812.09449.pdf> и другим источникам, часто используемыми методами в наиболее производительных решениях являются:

* на уровне символов – контекст, полученный с помощью LSTM, CNN, LSTM-CNN
* на уровне слов – эмбеддинги word2Vec, GloVe
* гибридные – словари, POS-тегирование, контекстуальные эмбеддинги ELMo и BERT

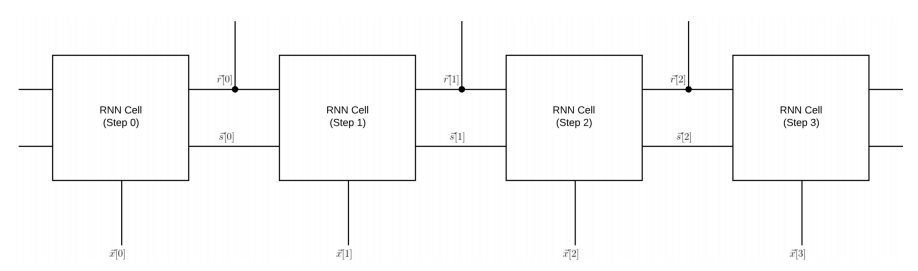
Для второго этапа наиболее популярными алгоритмом является LSTM(точность до 93.5% в комбинации с декодером CRF). Также высокими показателями точности обладают CNN и Transformer.

Наиболее распространенным алгоритмом декодера среди точных решений является CRF. Также часто употребима фукнция softmax.

Таким образом, в ходе работы могут быть рассмотрены указанные алгоритмы и их модификации в качестве перспективных. Для достижения большей точности, при извлечении сущностей предлагается использовать:

* для векторного представления входных данных – POS - тегирование и эмбеддинги BERT в силу наличия общеупотребимых слов. Также предполагается, что точность будет повышена за счет использования словарей, так как в НДЭ большое количество профессионализмов.
* Для кодирования контекста выбран LSTM как наиболее популярный и имеющий модификации(biLSTM, LSTM+attention).
* Для декодирования тегов из контекста выбран CRF – также, согласно масштабу используемости. Так как CRF трудоемок в обучении при увеличении количества распознаваемых сущностей, на данном уровне возможно также использование LSTM или softmax.

Алгоритм долгой краткосрочной памяти(LSTM) – модификация рекуррентной нейронной сети(RNN), основная идея которой – повторное использование одних и тех же весов при последовательно обрабатываемых ею эмбеддингов слов. Развертка переиспользования «ячейки» RNN во времени:



В процессе работы модели, рассчитываются следующие параметры нейросети:



Здесь вектор – скрытое состояние, свой вклад в которое, в зависимости от весов модели, вносит каждое входное слово последовательности. На текущем шаге работы модели, состояние модели определяется на основании предыдущего выходного состояния ) и входного вектора, умноженные на соответствующие им веса. Функция G здесь – обычно гиперболический тангенс.

У RNN существует ряд проблем: «взрывные» и затухающие градиенты, проблема долгосрочных зависимостей. Данные недостатки отчасти решены в LSTM и его модификациях.