УДК 004.85

ИЗВЛЕЧЕНИЕ СТРУКТУРИРОВАННЫХ ЗНАНИЙ ИЗ БИОМЕДИЦИНСКИХ СТАТЕЙ

Сазанович В.В. (Университет ИТМО)
Научный руководитель – кандидат технических наук, доцент Степанов Д.В.
(Университет ИТМО)

Профессиональные биологи создают базы знаний для представления сущностей и связей между ними. Такие базы полезны в различных областях биомедицины, однако процесс их создания трудоемкий и дорогостоящий. Более того, число публикаций в области биомедицины растет экспоненциально, и ручное извлечение знаний становится труднее. В данной работе предлагается инструмент, извлекающий связи между сущностями, используя методы глубокого обучения. Представленный алгоритм хорошо масштабируется и может извлечь и обработать информацию из всех опубликованных статьей. Такой подход может быть использован для приоретизации чтения статей и ускорения создания баз знаний.

Введение. На данный момент существует огромное количество биомедицинских баз, представлющих структурированные знания. В данной работе, нас интересуют базы, содержащие такие сущности как гены, болезни, вещества, а также связи между ними. Построение алгоритмов для извлечения такой информации является активной областью исследований, и на данный момент существует множество различных архитектур. Эта работа является первой работой, в которой алгоритм извлечения связей был применен к корпусу всех опубликованных биомедицинских статей.

Основная часть. Основой данной работы является алгоритм основанный на предобученной модели – BERT. Данная модель строит контекстуальные вложения входных токенов, что позволяет использовать ее для задач обработки естественного языка. Контекстуальные вложения затем с помощью механизма внимания и операции MaxPool преобразуются в вероятности наличия связи. Использование предобученной модели и механизма внимания, позволило получить результаты сопоставимые с текущими в области извлечения связей на произвольных текстах, а также превзойти предыдущие работы в области биомедицины.

Базой для обучения является “Comparative Toxicogenomics Database” (CTD). Для получения текстов статей и обогащения связей из CTD используется готовый веб инструмент – PubTator Central. PubTator предлагает автоматическую разметку генов, болезней, веществ, мутаций, организмов и линий клеток. Эта разметка и используется в нашем подходе. Следует отметить, что в дальнейшем планируется отказаться от готовой разметки и использовать собственную систему.

Представленный алгоритм справляется с извлечением связей из 30 миллионов статей в течение одного дня, используя параллельные вычисления на 8 графических процессорах.

Выводы. Представленный алгоритм извлечения связей достигает сопоставимых с известными подходами результатов на общих текстах, и превосходит предыдущие работы в области биомедицины. Представленный инструмент справляется с разметкой всех опубликованных биомедицинских статей в течение одного дня. Результатом работы является база извлеченных связей.

Сазанович В.В. (автор) Подпись

Степанов Д.В. (научный руководитель) Подпись