Автоматические методы извлечения сущностей из биомедицинских текстов

Кабанов Андрей Вячеславович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: arshehremen@gmail.com

Научный руководитель *— Лукашевич Наталья Валентиновна*

В области автоматического анализа и распознавания текста на естественных языках значимую роль занимает направление изучения биомедицинских текстов. Оно имеет широкое практическое применение, как своевременное извлечение важной информации в области биомедицины и здравоохранения. Для задачи извлечения сущностей из биомедицинских текстов активно размечаются и разрабатываются новые наборы данных на разных языках мира, в том числе на русском.

В данной работе представлено описание совершенно нового набора данных NEREL BIO, разрабатываемого НИВЦ МГУ. Набор данных предназначен для обучения моделей извлечения биомедицинских сущностей. Данные размечены аннотациями в формате BRAT [1]. Аннотации, созданные в BRAT, связаны с определенными отрезками текста с помощью смещения символов. Каждой аннотации соответствует идентификатор и структура разметки, которая зависит от её типа. Набор данных имеет иерархическую разметку, с максимальным уровнем вложенности равным 5. Технически это усложняет работу с набором данных, вследствие чего используются более сложные модели машинного обучения, требующие больших вычислительных ресурсов.

В работе представлены методы для более эффективного обучения модели, что значительно уменьшило число потребляемых вычислительных ресурсов.

Набор данных NEREL BIO на данный момент содержит 467 размеченных текстов, суммарно размечено 37226 сущностей, выделено 46 классов. Предварительно было выбрано 26 классов, в каждом из которых не менее 160 примеров. Исключенные классы, в основном, совсем не относились к биомедицине. Кроме этого для ряда экспериментов было выбрано 4 класса с распределённой частотностью.

В качестве базовой модели машинного обучения была выбрана модель MRC [2] на основе BERT [3]. MRC - Machine Reading Comprehension, модель, для задачи ответа на вопрос, где вопросом

является содержательный текст и описание некоторой сущности, а ответом - позиции этой сущности в тексте. Например, для поиска сущности “болезнь” на вход модели подаётся текст:

«Болезнь, нарушение нормальной жизнедеятельности организма, возникающее в ответ на действие патогенных факторов. Обследованы 63 пациента с хронической мигренью».

На выходе модели ожидается посимвольная позиция словосочетания “хронической мигренью” в тексте *“Обследованы 63 пациента с хронической мигренью*”.

В числе экспериментов представлены методы для эффективной работы с большим набором данных высокой сложности, как NEREL BIO. Они позволяют быстро и с меньшим числом вычислительных ресурсов достичь лучших оценок вычислений. Среди таких методов:

1. обучение модели на наборе данных по частям для каждого класса отдельно;
2. предварительная бинарная классификация для фильтрации и обучение оригинальной модели только на целевых, положительных примерах;
3. сокращение набора данных от очень похожих друг на друга примеров.

В работе проведен анализ набора данных NEREL-BIO. На основе современных моделей автоматической обработки текстов выполнены эксперименты, подтверждающие качество и репрезентативность нового набора данных.

Литература

1. Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., Tsujii, J. brat: a Web-based Tool for NLP-Assisted Text Annotation. EACL, 2012, pages 102–107.
2. Zhang, Z., Zhao, H., Wang, R. Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond // ArXiv, abs/2005.06249, 2020. 5
3. Devlin, J., Chang, M., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pages 4171–4186.