**УДК 004.85**

# РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ В БИОМЕДИЦИНСКИХ СТАТЬЯХ

**Утикеев С.М.** (Университет ИТМО)
**Научный руководитель – кандидат технических наук, доцент Степанов Д.В.**
(Университет ИТМО)

Биологам в профессиональной сфере для более удобной работы с сущностями и связями между ними приходится создавать определённые базы данных, агрегирующие имеющуюся предметную информацию. Создание таких баз является трудоёмким и дорогостоящим процессом, так как производится зачастую вручную. Данная работа предлагает инструмент, который позволит размечать биомедицинские тексты с помощью методов глубокого обучения. Представленная модель может использоваться в качестве вспомогательного блока в моделях, решающих другие задачи из области, такие как информационный поиск и извлечение связей между сущностями.

**Введение.** На данный момент область биомедицины включает в себя много разделов, по каждому из которых есть множество статей, число которых растёт с каждым днём. В текущей ситуации биологи вынуждены создавать различные базы знаний, чтобы извлекать наиболее важные понятия, сущности и отрывки информации вручную. Подход с применением технологий машинного обучения позволит значительно ускорить данный процесс, тем самым решая проблему экспоненциального роста числа статей, необходимых в разметке. Предлагается разработать инструмент для разметки текстов из предопределённого набора сущностей. Подобный инструмент помимо наглядного представления сущностей в статьях может помочь в улучшении результатов других задач из области, таких как извлечение связей и информационного поиска.

**Основная часть.** В качестве основной модели предлагается взять модель BERT, предобученную на корпусе биомедицинских текстов, также известную как BioBERT. После этого, обратившись к ряду датасетов, предоставляющих разметку генов, заболеваний, таксонов и других сущностей, можно построить разметчики отдельных конкретных сущностей. Так как разметчики одной сущности могут конфликтовать друг с другом (например, разметчик генов и разметчик болезней пометили одно и то же слово как соответствующую сущность), предлагается построить дополнительный разборщик подобных конфликтов и обучать его параллельно с разметчиками и предобученной моделью BERT. Предполагается, что данный подход позволит повысить точность и за счёт использования более узконаправленных разметчиков, и за счёт определённого элемента модели, отвечающей исключительно за разрешение конфликтов.

**Выводы.** В рамках данной работы разработана идея разрешителя конфликтов для разметчиков единичных сущностей. В реализации находится разработка прототипа модели, который позволит проверить гипотезу о возможности применения данного подхода для повышения качества разметки биомедицинских текстов.

Утикеев С.М. (автор) Подпись

Степанов Д.В. (научный руководитель) Подпись