Автоматизированное построение спецификации эксперимента

Бондарев Даниил Борисович

Студент

Физтех-школа ПМИ МФТИ, Долгопрудный, Россия

E-mail: bondarev.db@phystech.edu

Научный руководитель *— Хританков Антон Сергеевич*

Для воспроизводимости и анализа эксперимента в области искусственного интеллекта необходимо иметь подробную и актуальную спецификацию. Так, например, содержание статьи, публикуемой на конференции NeurIPS, должно соответствовать требованиям о воспроизводимости эксперимента (The Machine Learning Reproducibility Checklist). Создание спецификации является рутинной и ресурсозатратной задачей, особенно в случае сложного, долгосрочного, развивающегося со временем проекта. Таким образом, возникает необходимость в автоматизации построения спецификации эксперимента.

Актуальность задачи обуславливается необходимостью воспроизведения результатов как в исследовательской деятельности, так и в задачах индустрии. Так, например, это позволит другим исследователям подтвердить полученные результаты и использовать их в своих научных работах.

Рассмотрим задачу автоматизации построения спецификации на основе репозитория, в котором содержутся, используемые в работе, научные статьи. Данная задача может быть декомпозирована на несколько подзадач:

1. Предобработка текстовых данных;
2. Извлечение сущностей из текста;
3. Стандартизация извлеченных сущностей;
4. Формирование спецификации.

Основная проблема заключается в наиболее точном извлечении сущностей из текста. Для решения данной задачи было рассмотрено применение вопросно-ответных систем (question-answering), а именно архитектура нейронной сети Longformer [1], основанная на Transformer [2]. Longformer позволяет обрабатывать последовательности большей длины, при сравнении с сетями глубокого обучения, основанных на Transformer. Данное свойство является необходимым

при извлечении информации из текста статьи, так как это позволит избежать потери данных.

Для апробации метода были выбраны несколько статей из области искусственного интеллекта, а также несколько сущностей для извлечения. Валидация результатов была проведена в ручном режиме, которая показала применимость данного метода. Так, например, для статьи [2] удалось, предварительно выделив секции статьи, извлечь информацию о размерах обучающей выборки, количестве итераций обучения, а также вычислительных комплектующих.

Таким образом, решение задачи автоматизированного построения спецификации возможно с использованием вопросно-ответных систем.

Литература

1. Vaswani A, Matthew E., Cohan A. 2020. Longformer: The long-document transformer. arXiv:1706.03762.
2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J, Jones L., Gomez A. N., Kaiser L, Polosukhin. 2017. Attention Is All you Need. In NIPS, 2017