УДК 004.85

Распознавание именованных сущностей в научных текстах с использованием архитектуры «трансформер»

Утикеев Станислав, Университет ИТМО

Соломатов Константин — руководитель проекта в JetBrains s.r.o.

**Научный руководитель — Степанов Д. В., к.т.н.,
программист в ООО ИнтеллиДжей Лабс**

Введение

Область обработки естественного языка — обширная область задач глубокого машинного обучения, изучающая компьютерный анализ и синтез естественных языков. Среди данных задач встречаются такие задачи как машинный перевод, генерация текста, его разметка, информационный поиск, аннотирование и упрощений частей текста, анализ тональности ответа и многие другие. Одной из таких задач, связанных непосредственно с анализом и аннотированием, является задача распознавания именованных сущностей.

Распознавание именованных сущностей в разрезе научных текстов позволит упростить их аннотирование, что приведёт к лучшему поиску статей, поиску релевантной информации, необходимой учёному-исследователю при проведении собственного исследования. Так как в статьях могут присутствовать неизвестные для модели наименования сущностей, в рамках работы предлагается рассмотреть ряд подходов при обучении, которые позволят не только регуляризовать модель, но и размечать сущности на основе вероятностного анализа семантического сходства.

Цель работы

Целью данной работы является реализация модели для поиска и разметки именованных сущностей в научных текстах. Также в рамках работы предлагается изучить различные модификации архитектуры «трансформер» с целью получения лучших результатов в области обработки естественного языка и конкретно задаче распознавания именованных сущностей.

Базовые положения исследования

Задача распознавания именованных сущностей в тексте является одной из подзадач извлечения информации из текстов. Данная задача возникает в сферах медицины и журналистики, анализе архивных документов, научных статей, записей в блогах. В задачах обработки естественного языка за последнее время лучшие результаты показывают нейронные сети, использующие механизм внимания. Данный механизм позволил перейти от рекуррентных сетей к новой архитектуре — «трансформер» [2].

Эта архитектура позволила не только достичь state-of-the-art результатов во многих задачах обработки естественного языка, но и ускорить время обучения модели. При этом на сегодняшний день появляются всё новые архитектуры на основе «трансформера» [1, 3, 4], которые упрощают исходную модель, что позволяет работать с большими объёмами текста и позволяют лучше бороться с проблемами переобучения.

Для решения ряда проблем, связанных с переобучением модели, предлагается изучить возможности комбинирования различных подходов [1, 3, 5] к обучению нейронных сетей, выбрать те из них, что позволяют достичь лучших оценок на корпусах научных текстов, имеющих довольно специфическую структуру и лексику, отличную от обычных художественных и публицистических текстов.

Результаты

В рамках данной работы реализована базовая модель «трансформер» из статьи [3].

На данный момент ведётся исследование применения различных механизмов к улучшению качества работы модели. В реализации также находятся базовые модели, взятые за основу для оценки качества, из статей [4, 6].

Соломатов Константин: реализация базовой модели «трансформер» и построение архитектуры проекта.

Степанов Денис: идеи модификаций модели и реализация моделей на корпусах научных текстов.

Утикеев Станислав: реализация baseline-моделей, взятых за основу для сравнения качества работы моделей, исследование применения механизмов к архитектуре «трансформер» в рамках улучшения качества модели.

Список литературы

1. Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. arXiv:1808.04444, 2018.

2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. CoRR, abs/1706.03762, 2017.

3. Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert:Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2018.

5. Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. CoRR, abs/1508.07909, 2015.

6. Kevin Clark, Minh-Thang Luong, Christopher D. Manning, Quoc V. Le. Semi-Supervised Sequence Modeling with Cross-View Training. CoRR, abs/1809.08370, 2018.