Секция «Дискретная математика и математическая кибернетика»

Выделение ключевых терминов в научных текстах

Научный руководитель – Батура Татьяна Викторовна

Березин С.А.[1], *Паульс А.Е.*[2]

1 - Новосибирский государственный университет, Механико-математический факультет, Новосибирск, Россия, *E-mail: s.berezin@g.nsu.ru*; 2 - Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия, *E-mail: a.pauls@g.nsu.ru*

Выделение именованных сущностей (NER) это задача выделения в тексте слов или их сочетаний, обозначающих объект или явление определённой категории, например, названия организаций, имена людей и т.д. На сегодняшний день, большая часть исследований в области выделения именованных сущностей посвящена решению этой задачи на текстах разговорного, художественного или официально-делового стилей речи. В этой же работе мы рассмотрели проблему выделения сущностей из научных тестов.
В качестве данных нами был использован корпус SciERC [1], описывающий такие типы сущностей как “task”, “method”,“material”,“other scientific term” и другие. Этот набор данных включает 500 аннотаций к научным работам, полученых из Semantic Scholar Corpus. В каждой из них выделенны ключевые сущности и взаимосвязи между ними. Все представленные аннотации являются частью научных статей посвященных технологиям искуственного интеллекта.

Будучи интуитивно понятной для людей, эта задача долгое время находились за пределами возможностей автоматизированных систем. Многие годы лучшие решения основывались на некоем наборе правил составленных вручную или автоматически. Мы же в своих экспериментах мы использовали языковые модели RoBERTa [2] и ELECTRA [3]. Разбив имеющиеся данные на обучающую и тестовую выборку в отношении 70/30 мы провели ряд экспериментов по обучению нейросетей. Лучший достигнутый нами результат F1 macro = 0.58, что говорит о сложности решения поставленной задачи, т.к. те же самые архитектуры на более классическом наборе данных CoNLL 2003 дают результата F1 macro = 0.97. В дальнейших исследованиях мы планируем улучшить целевую метрику с помощью аугментации обучающих текстовых данных, после чего полученное решение станет основой для ПО позволяющего индексировать научные статьи и extract информацию из них.

Источники и литература

1) 1. Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction Yi Luan, Luheng He, Mari Ostendorf and Hannaneh Hajishirzi. EMNLP, 2018
2) 2. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.
3) 3. Clark, Kevin, Minh-Thang Luong, Quoc V. Le and Christopher D. Manning. “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.” ArXiv abs/2003.10555 (2020): n. pag.