ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 104–105

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ruSciBERT: ЯЗЫКОВАЯ МОДЕЛЬ НА БАЗЕ АРХИТЕКТУРЫ ТРАНСФОРМЕР ДЛЯ ПОЛУЧЕНИЯ СЕМАНТИЧЕСКИХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ НАУЧНЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

© 2022 г. Н. А. Герасименко[1,*], А. С. Чернявский[1], М. А. Никифорова[1]

Представлено академиком РАН А.Л. Семеновым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Значительный рост числа научных публикаций и количества научных отчетов делает задачу их обработки и анализа сложной и трудозатратной. Языковые модели, основанные на архитектуре Трансформер и предобученные на больших текстовых коллекциях, позволяют качественно решать множество задач анализа текстовых данных. Для работы с научными текстами на английском языке существуют модели SciBERT [1] и ее модификация SPECTER [2], однако они не поддерживают русский язык в связи с малым количеством текстов в обучающей выборке. Кроме того, способ оценки качества языковых моделей для научных текстов, бенчмарк SciDocs, также поддерживает только английский язык. Предлагаемая модель ruSciBERT позволит решать широкий спектр задач, связанных с анализом научных текстов на русском языке, а прилагаемый к ней бенчмарк ruSciDocs позволит оценивать качество языковых моделей применительно к этим задачам.

Ключевые слова: языковая модель, семантические представления, SciBERT, SciDocs

DOI: 10.31857/S2686954322070074

SciBERT является языковой моделью, обученной на многодоменном корпусе научных статей, написанных преимущественно на английском языке. Авторы предложили взять базовую модель BERT и дообучить ее на задаче предсказания маскированных токенов. Результаты, полученые авторами на нескольких задачах классификации и NER для научных статей, значительно превосходят результаты базовой модели. Дополнительно обученный токенизатор позволил улучшить получаемое качество language modeling. Мы используем похожие идеи и предлагаем дообучение модели RoBERTa на русскоязычном корпусе научных текстов с собственным токенизатором. Данную модель мы называем RuSciBERT. В качестве базовой модели нами была выбрана RoBERTa в связи с тем, что она обучена на расширенном количестве данных, большем количестве задач и достигла лучших результатов по сравнению с базовым BERT.

SciDocs является бенчмарком для оценки качества семантических векторных представлений, получаемых с помощью языковых моделей. Он включает в себя 4 типа задач:

1. Классификация на основе классификаторов MAG и MeSH

2. Предсказание цитирования на основе Semantic Scholar Academic Graph

2.1. Прямые цитаты (задача ранжирования)

2.2. Социтируемые статьи (задача ранжирования)

3. Предсказание активности пользователей Semantic Scholar

3.1. Сопросматриваемые статьи (задача ранжирования)

3.2. Сопрочитываемые статьи (задача ранжирования)

4. Рекомендации статей, похожих на статью-запрос (задача ранжирования).

ruSciBERT планируется обучить на датасете, включающем около 1 млрд токенов. Данные для обучения собраны из открытых источников, позволяющих использовать данные для некоммерческих целей (например, из датасета Semantic Scholar Academic Graph). Размер словаря токенизатора в нашем случае равен 50265 по аналогии с базовой моделью RoBERTa.

Бенчмарк ruSciDocs планируется составить из задач, аналогичных части задач оригинального SciDocs:

1. Классификация на основе классификаторов

[1] *ПАО "Сбербанк", Москва, Россия*
*****E-mail: nikgerasimenko@gmail.com*

1.1. MAG – верхний уровень Microsoft Academic Graph, таксономии областей знания, составленной специалистами из Microsoft и Allen institute for AI

1.2. OECD из ЕГИСУ НИОКТР, государственного сайта для учета научно-исследовательских работ

2. Предсказание цитирования на основе данных из Semantic Scholar Academic Graph.

На данный момент мы обучили модель RuSciBERT на 300 млн токенов на двух эпохах. Она показывала хорошие результаты при заполнении пробелов в текстовых фразах, а также гораздо более низкий уровень перплексии на отложенной выборке по сравнению с общей языковой моделью ruBERT, обученной на текстах всех тематик. Так, RuSciBERT имеет перплексию 4.81, причем она монотонна снижается на последних шагах обучения, вследствие чего модель можно дообучать дальше. В то же время ruBERT имеет перплексию 9.64.

Примеры работы нашей модели заполнения маскированных токенов показаны ниже. В них маскированные токены обозначены через “<mask>”, а модель предсказывает 3 наиболее вероятных варианта токенов для замены.

1) “при использовании в усилителе мощности адаптивной измерительной <mask> появится возможность” -> 'системы', 'аппаратуры', 'станции'

2) “указанные оппоненты не имеют <mask> проектов и публикаций с соискателем” -> 'совместных', 'собственных', 'аналогичных'

3) “новый метод управления <mask> характеристиками ао фильтров” -> 'техническими', 'технологическими', 'функциональными'

RuBERT также показывает неплохие результаты, но некоторые из его вариантов заполнения являются менее удачными. Так, в первом примере среди предсказанного множество токенов есть 'технологии' (меньше подходит по смыслу чем остальные варианты), во втором – 'своих', а в третьем – 'всеми' (возможные варианты, но более общие, и поэтому менее качественные).

Основываясь на текущих промежуточных результатах, можно предположить, что ruSciBERT, обученный на датасете в 1 млрд токенов, покажет наилучшие разультаты на бенчмарке ruSciDocs по сравнению с другими существующими подходами.

СПИСОК ЛИТЕРАТУРЫ

1. Iz Beltagy and Kyle Lo and Arman Cohan. SciBERT: Pretrained Language Model for Scientific Text // EMNLP, 2019.
2. Arman Cohan and Sergey Feldman and Iz Beltagy and Doug Downey and Daniel S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers // ACL, 2020.