УДК 004.8

ПРИМЕНЕНИЕ ТРАНСФОРМЕРОВ ДЛЯ РАСПОЗНАВАНИЯ СУЩНОСТЕЙ В ТЕКСТАХ О ЦИФРОВОЙ ТРАНСФОРМАЦИИ ОБЩЕСТВА

Беген П.Н. (Университет ИТМО)
Научный руководитель – к.т.н. Митягин С.А.
(Университет ИТМО)

Рассмотрено применение подхода к задаче распознавания именованных сущностей в текстах по тематике цифровой трансформации общества на основе трансформеров. Представлен список сущностей, готовых для выявления.

Исследование направлено на решение проблемы кластеризации и классификации специализированных текстов по тематике цифровой трансформации общества в России. На данный момент существует проблема распознавания сущностей в таких текстах, поскольку имеющиеся наборы размеченных массивов русскоязычных текстов лежат вне области цифровой трансформации и существенно сужают корректность и точность тематической классификации для последующего анализа, прогнозирования и построения трендов развития.
В исследовании используется массив текстов новостных сообщений «Развитие цифрового государственного управления в Российской Федерации», сформированный Центром технологий электронного правительства за период 2010–2021 гг. и насчитывающий более 8,5 тыс. единиц. В текстах о цифровой трансформации общества выделены следующие сущности для распознавания: Уровень, Отрасль, Технология, Статус, Организация, Участники, География. Для некоторых сущностей задан определенный список значений, например для сущности «Уровень» список следующий: - муниципальный (т.е. местного значения); - региональный; - федеральный; - международный.
При изучении и анализе текста в рамках задачи распознавания именованных сущностей (NER) исследователи зачастую используют рекуррентные нейронные сети с LSTM-блоком либо традиционные методы машинного обучения, как Наивный Байесовский классификатор, SVM, деревья решений и др. С 2017 года начался период активного перехода к усовершенствованным моделям типа трансформер (Transformer), которые начали показывать более качественные результаты (BERT, GPT и др.).
Архитектура трансформера основана на глубоких нейронных сетях (deep learning) и состоит из кодировщика и декодировщика. Каждый кодировщик состоит из механизма самовнимания, self-attention, (вход из предыдущего слоя) и нейронной сети с прямой связью (вход из механизма самовнимания). Каждый декодировщик состоит из механизма самовнимания, механизма внимания к результатам кодирования и нейронной сети с прямой связью (вход из механизма внимания). Отличительной особенностью трансформеров является то, что трансформеры не требуют обработки последовательностей по порядку, в отличие от рекуррентных нейронных сетей, за счёт чего повышается качество результата. Крупные IT-компании, такие как Яндекс и Google, активно используют трансформеры в машинном лингвистическом переводе, в голосовых помощниках, обработке изображений и видео, при работе с текстами. Планируется, что использование трансформеров в текстах о цифровой трансформации общества позволит выявлять сущности с гораздо более высокой точностью, по сравнению с традиционными методами машинного обучения.
Работа выполнена в рамках проекта НИР Университета ИТМО № 621304, «Разработка сервиса тематической кластеризации корпуса текстов «Развитие цифрового государственного управления в Российской Федерации» на основе машинного обучения».