УДК 004.891.3

Лингвистическая проверка эссе полного цикла методами глубокого обучения

Д.Д. Астафуров МОУ «Лицей прикладных наук им. Д.И.Трубецкова» г.Саратов

В связи с увеличением количества письменных работ и необходимостью увеличения скорости их проверки была поставлена задача разработать алгоритм способный переводить изображение с работой в текст и далее осуществлять лингвистический анализ полученного документа. Были выбраны для анализа эссе по русскому языку, литературе, истории, обществознанию.

Проблема заключается в необходимости автоматизации проверки творческих работ, причем на этапе как оцифровывания, так и лингвистической проверки. Поставлены задачи: собрать данные, проанализировать датасет, решить задачу детекции текста, решить подзадачи лингвистического анализа на: ошибки в построении структуры, речевые ошибки, грамматические ошибки, логические ошибки, фактические ошибки, этические ошибки.

Были найдены датасеты для анализа: детекция слов - 684 тетрадных листа, транскрибирование слов - 66 тыс. слов, лингвистический анализ - 7 тыс. сочинений по русскому языку, литературе, истории, обществознанию.

Для задачи детекции слов и выделения масок была выбрана архитектура Mask R-CNN. В качестве метода борьбы с переобучением были добавлены аугментации, искажающие цвет и форму изображения, а также copy-paste.

Проблема транскрибирования текста была решена с помощью архитектуры CRNN (Convolutional Reccurent Neural Network) с использованием CTC Loss.

Модель сегментации текста основана на архитектуре BERT based on transformer. Были выделены следующие классы: ПРИМЕР, СЯП, ПОНЯТИЕ, ЛОГИКА, ТЕОРИЯ, ПОЯСНЕНИЕ, ПРИЧИНА, СЛЕДСТВИЕ, ИДЕЯ, ПРОБЛЕМА, ОТНОШЕНИЕ, ПОЗИЦИЯ, РОЛЬ, АРГУМЕНТ, СВЯЗЬ, ОЦЕНКА, ИСП.

Результаты работы в метриках: детекция слов – доля найденных слов = 0.89, транскрибирование – CER=0.032, грамматические ошибки – f1-мера=0.51, речевые ошибки – f1-мера=0.42.