УДК 004.8

МЕТОДЫ И ТЕХНОЛОГИИ МАСШТАБИРУЕМОСТИ АЛГОРИТМОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА МЕДИЦИНСКИХ ТЕКСТОВ

Шайкина А.А. (Университет ИТМО), **Функнер А.А.** (Университет ИТМО)
Научный руководитель – к.т.н., доцент Ковальчук С.В.
(Университет ИТМО)

Аннотация: Данная работа посвящена методам и технологиям масштабируемости алгоритмов интеллектуального анализа медицинских текстов. Рассматриваемые методы позволяют использовать опыт, накопленный при решении одной задачи, для решения другой, аналогичной проблемы, то есть дообучать имеющиеся модели на небольшом количестве новых данных вместо обучения новых.

Введение. Для моделей контролируемого обучения ученые обычно используют данные с метками, которые задаются вручную экспертами. Таким образом, при необходимости использовать имеющуюся модель для нового медицинского центра, существуют два способа: разметить данные этого центра и переобучить модель или использовать ранее подготовленную модель. Во втором случае экономится время, требуется значительно меньше данных. Целью данной работы является исследование методов и реализация масштабируемости алгоритмов анализа медицинских текстов с возможностью применения к новым данным моделей, обученных на данных другого корпуса.

Основная часть. В основе традиционных методов машинного обучения лежит предположение о том, что данные для обучения и данные для тестирования берутся из одного источника. Однако в некоторых сценариях данное предположение неверно, так как требуемые для обучения данные могут быть дорогостоящими или их получение затруднительным. Поэтому в данной работе рассматривается вопрос использования готовых, уже обученных моделей для уточнения результатов новых моделей, построенных для медицинских учреждений со схожей направленностью. На данный момент для внедрения алгоритмов обработки медицинских текстов в новое медицинское учреждение необходимо заново настраивать модели на данных этого учреждения. Рассматриваемые модули обработки медицинского текста в своей основе используют различные модели, поэтому требуют различного решения задачи масштабируемости, однако для этого используется общий принцип переноса обучения (Transfer learning). Например, применение данного принципа для модуля тематической сегментации реализовано так, что для тематических моделей на новых данных задаётся родительская модель, уже обученная ранее на схожих данных, результаты которой учитываются с весом, указанным при инициализации модели.

Выводы. В данной работе исследованы вопросы масштабируемости алгоритмов интеллектуального анализа тексов применительно к медицинским слабоструктурированным данным. В рассматриваемой задаче тематической сегментации наборов анамнезов двух медицинских центров перенос обучения реализован с применением к каждому набору моделей, включающих родительские – обученные на другом для него корпусе.

Шайкина А.А. (автор) Подпись

Ковальчук С.В. (научный руководитель) Подпись