**УДК 004.912**

# АНАЛИЗ ПОДХОДОВ К ОБРАБОТКЕ И АНАЛИЗУ ТЕКСТОВ, СОДЕРЖАЩИХ МАТЕМАТИЧЕСКИЕ ВЫРАЖЕНИЯ

**Рыбин А.С.** (Университет ИТМО)

**Научный руководитель – к.т.н. Махныткина О.В.**

(Университет ИТМО)

В работе проведён сравнительный анализ существующих подходов к автоматической обработке математических формул и построения их признакового описания.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №620183 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе моделей-трансформаторов и понимания естественного и математического языка».

Математические выражения (формулы) являются важной частью научных документов (например, научных статей), однако их автоматический анализ вызывает затруднения, т.к. требует обработки одновременно как естественного языка, так и языка математических формул. В то время как в области *NLP* (обработки естественного языка) уже устоялись подходы, основанные на машинном обучении, их применение в области *MLP* (обработки математического языка) пока только изучается.

Извлечение признаков важный шаг при решении любой задачи с применением методов машинного обучения. Признаковое описание математических формул, может быть, использовано для решения задач поиска похожих формул, извлечения структурированной информации из формул и перевода из одного представления формулы в другое.

Для обучения и тестирования моделей в открытом доступе находятся несколько датасетов. *Wikipedia* на 30 июля 2014 содержит порядка $2 \cdot 10^5$ формул в формате *LaTeX*, которые содержат минимум 2 переменных и 3 оператора. Датасеты *SigMathLing arXMLiv* и *NTCIR 11/12 MathIR arXiv* содержит порядка $1.2 \cdot 10^6$ и $1 \cdot 10^5$ HTML документов с формулами в формате *MathML* соответственно.

В работе был представлен обзор существующих подходов построения признакового описания объектов в областях NLP и MLP для использования методов машинного обучения и, в частности, методов глубокого обучения. Рассмотрены 4 группы методов: мешок слов, дистрибутивная семантика, универсальная языковая модель и Equation Embeddings. Для сравнения методов были установлены критерии, которые отражают важные аспекты при решении задачи MLP (NLP). В ходе обзора и сравнения было установлено, что наиболее перспективным выглядит адаптация метода *Doc2Vec* на символьном уровне в силу того, что данный метод прост в реализации, не требует для обучения корпус большого объема и вычислительных ресурсов, но при этом обладает достаточной репрезентативностью что подтверждается исследованиями.