**УДК 004.89**

# ИЗВЛЕЧЕНИЕ ЗНАНИЙ ОБ УЧАСТНИКАХ ДИАЛОГА С ПРИМЕНЕНИЕМ МОДЕЛЕЙ ТРАНСФОРМЕРОВ

**Скрыльников С.С.** (Университет ИТМО)
**Научный руководитель – к. т. н. Махныткина О.В.**
(Университет ИТМО)

**Введение.** В данной работе рассматривается применение трансформер моделей для решения задачи извлечения знаний из диалоговых данных. Представлено описание используемого набора данных, их предобработка и описание архитектуры применяемой сети трансформера. Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №622282 «Разработка русскоязычного персонифицированного диалогового агента с динамической долгосрочной памятью».

**Основная часть.** Извлечение знаний из текстов на естественном языке становится всё более применимым в высокоуровневых задачах обработки естественного языка для получения формализованных знаний о тексте. С целью построения диалогового агента, который будет иметь представление о собеседнике, необходимо решение задачи извлечения информации о нём в процессе диалога[1].
Для проведения исследований был выбран датасет Toloka Persona-chat на русском языке, состоящий из диалогов разговорного жанра и описанием персон, принимающих участие в беседе. Общий размер набора данных составляет 10013 диалогов и 1505 уникальных персон.
Для обучения и тестирования моделей, данный датасет был случайным образом разбит на обучающую выборку, включающую в себя 9 018 диалогов или 152 174 реплик и тестовую, включающую 995 диалогов или 16 609 реплик. Тестовая выборка в ходе подготовки данных для соответствия задаче извлечения знаний была размечена людьми и имеет необходимую эталонную разметку для дальнейшего обучения t5 модели для автоматического поиска знаний[2].

**Выводы.** В области машинной обработки естественного языка, в последнее время наиболее используемыми стали модели трансформеры вследствие их направленности на обработку последовательностей данных[3]. Представленная в работе задача является text-to-text преобразованием, вследствие чего, для практического решения была выбрана трансформер модель Text-To-Text-Transfer-Transformer (T5). В качестве метрики оценки точность модели выбрана метрика BLEU. В данной работе рассматривается применимость данной архитектуры к задаче извлечения знаний об участниках диалога из данных на естественном языке.

**Список использованных источников**:

1. Zhang S. et al. Personalizing dialogue agents: I have a dog, do you have pets too? //arXiv preprint arXiv:1801.07243. – 2018.
2. Zhong P. et al. Towards persona-based empathetic conversational models //arXiv preprint arXiv:2004.12316. – 2020
3. Thulke D. et al. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog //arXiv preprint arXiv:2102.04643. – 2021.