УДК 004.89

АНАЛИЗ МЕТОДОВ АУГМЕНТАЦИИ ТЕКСТОВЫХ ДАННЫХ С СОХРАНЕНИЕМ ОТЛИЧИТЕЛЬНЫХ ХАРАКТЕРИСТИК РЕЧИ ЧЕЛОВЕКА

Матвеева А.А.
(Университет ИТМО)
Научный руководитель — к.т.н. Махныткина О.В.
(Университет ИТМО)

В данной работе рассмотрены методы текстовой аугментации (LAMBADA, Paraphrases Syntax Tree, Context Augmentation), способных обеспечить сохранение отличительных характеристик письменной речи человека.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №621296 «Разработка технологий для персонификации разговорного искусственного интеллекта».

В настоящее время особо востребована разработка различного рода диалоговых помощников. Отдельной задачей при их разработке является персонификация данных диалоговых помощников для повышения лояльности и вовлеченности в беседу пользователей, что может являться конкурентным преимуществом для компаний их использующих. Главным ограничением при персонификации диалоговых помощников является отсутствие больших наборов данных с диалогами, содержащих характеристики персон.

Анализ методов аугментации проводится на базе персонифицированных наборов данных «Toloka ru PersonaChat» и «Persona Chat». «Toloka Persona Chat Rus» - набор данных на русском языке из 10 013 диалогов с 1505 различными персонами, описанными 5 предложениями вида «Я рисую», «Я живу за границей». «Persona Chat» - англоязычный набор данных, состоящий из 10 907 диалогов и 1 155 персон, каждая из которых описана 3-5 предложениями.

В данной работе рассматривается возможность применения методов аугментации (LAMBADA, Paraphrases Syntax Tree, Context Augmentation) для решения задачи аугментации диалогов и описаний персон с сохранением отличительных характеристик письменной речи отдельно взятой персоны. Также был проведен сравнительный анализ рассмотренных методов аугментации.

- LAMBADA – метод аугментации на основе предобученной на больших корпусах языковой модели GPT с фильтрацией аугментированных данных с помощью BERT классификатора;
- Paraphrases Syntax Tree – метод аугментации, заключающийся в создании из исходного предложения дерева зависимостей и последующего его преобразования в соответствии с грамматикой. Преобразованное дерево зависимостей используется для создания нового выражения;
- Context Augmentation – метод дополнения данных заменой слов другими словами, предсказываемыми с помощью двунаправленной языковой модели на основе контекстного окружения.