УДК 004.89

РАЗРАБОТКА МЕТОДА ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ О ПЕРСОНЕ ИЗ ДИАЛОГОВЫХ ДАННЫХ

Рудалева Е.А (Университет ИТМО)
Научный руководитель –к.т.н. Махныткина О.В
(Университет ИТМО)

Введение. В настоящее время широкую популярность в области речевых технологий набирают разговорные диалоговые системы (*chitchat bots*). Извлечение информации о персоне пользователя в данном направлении является широко исследуемой темой, поскольку позволяет адаптировать систему к конкретному лицу и собрать базу предпочтений и черт характера для генерации более естественных, приближенных к человеческим ответов. Данная работа представляет собой подбор методов извлечения информации о персоне пользователя из диалоговых данных с использованием лингвистических признаков.

Исследование было выполнено за счет финансирования университета ИТМО в рамках НИР № 622282 «Разработка русскоязычного персонифицированного диалогового агента с динамической долгосрочной памятью».

Основная часть. Работа посвящена разработке лингвистического правила, оптимально извлекающего информацию о персоне. В качестве основных задач были определены две: детектирование реплик, содержащих информацию о персоне, через бинарную классификацию и сопоставление данных персонифицированных реплик с имеющимися описаниями к конкретной персоне с помощью предобученной *sentence-transformers* [1] модель *sbert_synonymy* [2]. Подбор признаков для извлечения информации производился на синтаксическом и морфологическом языковых уровнях. Исследование проводится на размеченных вручную на русскоязычных данных *Toloka PersonaChat*, в которых репликам, содержащим информацию о персоне, присваивались метки, ассоциированные с готовыми описаниями персон.

Выводы. На основе рассмотрения известных и общедоступных баз диалоговых данных, а также существующих методов извлечения информации о персоне были сформированы комбинации возможных лингвистических правил, из которых на основе бинарной классификации было отобрано правило, основанное на адаптированной комбинации морфологических и лексических признаков. Результаты отработки алгоритма на предварительно размеченных данных смогли достичь значения метрик, близких к значениям сходных метрик для алгоритмов, основанных на случайных (естественных) англоязычных данных и нейронных сетях [3, 4].

Список использованных источников.

1. Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks //arXiv preprint arXiv:1908.10084. – 2019.
2. Детектор синонимичности фраз. Sbert_synonymy [Электронный ресурс]: https://huggingface.co/inkoziev/sbert_synonymy
3. Zhang, Saizheng, et al. "Personalizing dialogue agents: I have a dog, do you have pets too?." arXiv preprint arXiv:1801.07243 (2018)
4. Zheng Y. et al. A pre-training based personalized dialogue generation model with persona-sparse data //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 05. – С. 9693-9700.