**САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО**

**Институт компьютерных наук и технологий**

УТВЕРЖДАЮ

Руководитель ОП

В.Г. Пак

« » 2022 г.

ЗАДАНИЕ

на выполнение выпускной квалификационной работы

студенту Фурману Владиславу Константиновичу, гр. 3540203/00101

1. Тема работы: «Разработка и исследование системы автоматического упрощения текстов на японском языке».

2. Срок сдачи студентом законченной работы: 05.2022.

3. Исходные данные по работе:

3.1. Существующие решения в области упрощения текстов на японском языке:

* Attention Is All You Need / A. Vaswani [и др.]. — 2017. — arX iv: 1706.03762 [cs. CL].
* Goto I., Tanaka H., Machine Translation of Japanese News Simplification: Task Design, Data Set Construction, and Analysis of Simplified Text // Machine Translation Summit XV. — 2015. — Т. 1. — С. 17—31.

3.2. Библиотеки для разработки модели упрощения:

* PyTorch. Python machine learning framework. — URL: https://pytorch.org (дата обращения: 06.02.2022).
* MeCab. Japanese language tokenization. — URL: https://github.com/taku910/mecab (дата обращения: 21.05.2021).

3.3. Корпус упрощённых японских предложений: Katsuta A., Yamamoto K. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — Miyazaki, Japan: European Language Resources Association (ELRA), 2018. — URL: https://www.aclweb.org/anthology/L18 — 1072.

4. Содержание работы (перечень подлежащих разработке вопросов):

4.1. Особенности японского языка: письменность, морфология, грамматика, а также упрощение этих составляющих языка.

4.2. Существующие решения и наборы данных, обзор статей о модели Transformer, об улучшении упрощения увеличением корпуса с обучением без учителя, об упрощении новостей, а также о JSSS-корпусе и о корпусе SNOW T15.

4.3. Обзор использования упрощения в реальном мире: сервисы NHK News Web Easy, Simple English Wikipedia, а также описание текущего положения дел в упрощении текстов на естественных языках.

4.5. Теоретическая часть работы: описание этапов обработки естественного языка, существующих решений в области упрощения, а также подробный теоретический обзор модели Transformer.

4.6. Программная реализация системы, описание её деталей (модель Transformer, сервер и пользовательское приложение).

4.7. Апробация и тестирование реализованной системы упрощения на предложениях на японском языке, сравнение полученных результатов с существующими на сегодняшний день открытыми решениями.

4.8. Выводы и заключение.

5. Перечень графического материала (с указанием обязательных чертежей):

5.1. Архитектура модели Transfomer из оригинальной статьи («Attention Is All You Need»).

5.2. Скриншот из пользовательского приложения, разработанного в ходе данной ВКР.

6. Консультанты по работе: нет.

7. Дата выдачи задания 03.02.2020.

Руководитель ВКР \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Задание принял к исполнению 03.02.2020.

Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_