Министерство Науки и Высшего Образования Российской Федерации

Национальный Исследовательский Институт Высшая Школа Экономики

Факультет Компьютерных Наук

Школа Анализа Данных и Искусственного Интеллекта

РЕФЕРИРОВАНИЕ ТЕКСТА НА РУССКОМ ЯЗЫКЕ

Компьютерная лингвистика и анализ текстов

Студент *М.Д. Кирдин*

Преподаватель *Е.И. Большакова*

Москва, 2025г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Автоматическое реферирование текста является одной из основополагающих задач обработки естественного языка наряду с машинным переводом и распознаванием сущностей. Способы решения этой задачи делятся на две категории: извлекающие и генерирующие. Целью данной работы было провести сравнение этих подходов к решению задачи аннотирования текста. Извлекающие подходы были представлены алгоритмом *TextRank*, а генерирующие подходы — моделями с трасформерной архитектурой *FRED-T5-Summarize*, а также *rut5-base* с параметрами отрегулированными для решения задачи реферирования текстов на русском языке.

1 ХОД РАБОТЫ

Для сравнения двузх рахличных подходов было решено использовать специализированный датасет, предолженный Ахметгареевой А и др. [1]. Он состоит из 197 тыс. текстов в части предназначенной для обучения и 258 текстов поверенных вручную в части для тестов.

1.1 Реализация извлекающего алгоритма

Алгоритм *TextRank* является модификацией алгоритма *PageRank*, предложенного *Google* в 1998 году. В данной работе используется вариант данного алогоритма для извлечения предложений. Он основан на построении графа при помощи алгоритма *PageRank*, в котором вершинами являются предложения в тексте и извлечении n вершин с наибольшим значением внутренней метрики. В рамках данной работы был написан скрипт на языке *Python* с его реализацией.

Для построения графа необходима матрица сходств предложений в реферируемом тексте. Она была получена как набор попарных косинусных расстояний между суммами эмбеддингов отдельных токенов. Эмбеддинги и токенизатор были взяты из библотеки *SpaCy*.

1.2 Реализация генерирующих алгоритмов

Был написан скрипт на языке *Python*, который

РЕЗУЛЬТАТЫ

ЗАКЛЮЧЕНИЕ

СПИСОК ИСТОЧНИКОВ

1. Akhmetgareeva A., Kuleshov I., Leschuk V., Abramov A., Fenogenova A., Towards Russian Summarization: can architecture solve data limitations problems? // https://sberlabs.com/publications?publication=1600 (2024).