МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

СРАВНЕНИЕ МЕТОДОВ СУММАРИЗАЦИИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Атеняева Дмитрия Сергеевича

Научный руководитель
Старший преподаватель Е.Е.Лапшева

Зав. кафедрой
к.ф.-м.н., доцент М.В.Огнева

Саратов 2025

ВВЕДЕНИЕ

Актуальность темы. Одним из ключевых направлений в области обработки естественного языка является автоматическая суммаризация – процесс сокращения текста с сохранением его основного смысла. Такая технология позволяет существенно сократить время восприятия информации и повысить эффективность принятия решений, что особенно важно в новостной аналитике, юридических документах, образовательных материалах и других сферах, связанных с большими объёмами текстов.

Актуальность задачи автоматической суммаризации подтверждается активным развитием научных исследований и широким внедрением в прикладные решения. Суммаризация широко используется в интернет-браузерах, например, в Яндекс Браузере; виртуальных помощниках, чат-ботах, интеллектуальных поисковых системах и системах поддержки принятия решений [1]. В отчёте аналитической компании Gartner за 2022 год автоматическая генерация и резюмирование текстов названы одними из ключевых направлений в развитии корпоративных ИИ-систем [2].

С развитием трансформерных архитектур и появлением масштабных языковых моделей (например, BART, T5, GPT) существенно повысилось качество абстрактивных решений, что сделало их особенно перспективными для автоматизации обработки естественного языка.

Несмотря на наличие промышленных решений, использующих методы абстрактивной суммаризации для русского языка таких как Яндекс.Браузер с режимом пересказа содержания веб-страниц и модели GigaChat от Сбера – задача остаётся недостаточно проработанной в академической среде. Это связано не только с ограниченным числом открытых датасетов и языковыми особенностями, но и с ограничениями на использование самих моделей: большинство доступных решений предоставляют лишь ограниченное количество запросов, требуют значительных вычислительных ресурсов или

недоступны для локального развертывания, что затрудняет проведение масштабных экспериментов и воспроизводимых исследований [3].

Таким образом, существует потребность в комплексном анализе существующих решений, адаптированных к русскоязычным данным, с целью выявления наиболее эффективных подходов и оценки их применимости в различных прикладных задачах.

Целью бакалаврской работы является проведение сравнительного анализа абстрактивных методов суммаризации текстов на русском языке.

Для достижения поставленной цели в работе решаются **следующие задачи:**

1. Проанализировать существующие подходы и модели абстрактивной суммаризации текстов.
2. Изучить метрики оценки качества суммаризации.
3. Настроить вычислительную среду для практической части работы с использованием облачных ресурсов (Yandex Cloud).
4. Найти и подготовить датасет на русском языке для задачи суммаризации текстов.
5. Выполнить тематическую классификацию текстов для повышения эффективности обучения моделей.
6. Выбрать подходящие предобученные генеративные модели и дообучить их на выбранных данных.
7. Провести сравнение моделей по качеству суммаризации с применением выбранных метрик.
8. Проанализировать полученные результаты и сформулировать выводы о применимости исследованных методов.

Методологические основы исследования в области суммаризации текстов представлены в работах Альбины Ахметгареевой, Александра Абрамова, Ильи Кулешова, Влада Лещука, Алёны Феногеновой [3], Аман Кедиа, Майанк Расу [4], Раффел Колин [5].

Практическая значимость бакалаврской работы заключается в проведении сравнительного анализа методов абстрактивной суммаризации текстов на русском языке с учётом их тематической принадлежности.

Структура и объём работы. Бакалаврская работа состоит из введения, 3 разделов, заключения, списка использованных источников и 4 приложений. Общий объем работы – 74 страницы, из них 55 страниц – основное содержание, включая 51 рисунок и 1 таблицу, список использованных источников информации – 22 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические основы абстрактивной суммаризации текстов и методов её оценки» посвящён обзору современных подходов к автоматическому созданию кратких пересказов.

Абстрактивная суммаризация представляет собой задачу генерации нового текста, отражающего основные идеи исходного документа, но при этом не копирующего его дословно. В отличие от экстрактивной, где берутся фрагменты оригинала, абстрактивная суммаризация требует понимания смысла и генерации новых формулировок. Полученная аннотация должна быть грамматически корректной, логически связной и содержать только релевантную информацию, важность которой определяется контекстом и предметной областью.

Вначале рассматривается базовая архитектура sequence-to-sequence (Seq2Seq), в которой используется пара моделей: кодировщик и декодировщик. Кодировщик преобразует входной текст в векторное представление фиксированной длины, а декодировщик на его основе генерирует выходной текст. Такой подход нашёл широкое применение в задачах машинного перевода, суммаризации и генерации текста.

Затем внимание уделено трансформерным архитектурам, заменившим рекуррентные сети благодаря своей способности параллельно обрабатывать последовательности и эффективно учитывать контекст с помощью механизма внимания. Особый акцент сделан на моделях BART и T5.

BART (Bidirectional and Auto-Regressive Transformers) сочетает в себе автоэнкодерную архитектуру с автогенерацией, что позволяет ей эффективно решать задачи генерации текста, включая суммаризацию. Модель сначала «корраптит» (искажает) входной текст, а затем восстанавливает его, обучаясь на реконструкции исходных данных.

T5 (Text-To-Text Transfer Transformer) — универсальная модель, переводящая любые задачи обработки естественного языка в формат «текст → текст». Её ключевая особенность — единый подход к обучению: как

классификация, так и генерация формулируются как задачи преобразования строки.

Особое внимание в разделе уделено оценке качества моделей суммаризации. Несмотря на то что метрики BLEU, METEOR и chrF изначально разрабатывались для машинного перевода, они успешно применяются и в задачах генерации кратких текстов. Метрики ROUGE-n и ROUGE-L, напротив, изначально ориентированы на оценку суммаризации, сравнивая n-граммы, префиксы и предложения оригинала и сгенерированного текста. В совокупности, эти метрики обеспечивают как количественную, так и структурную оценку качества итогового резюме.

Второй раздел «Инструменты программной реализации» описывает программные средства, используемые в рамках практической части исследования.

Рабочая среда была развернута на облачной платформе Yandex Cloud, что позволило преодолеть ограничения по ресурсам, часто возникающие при работе с глубокими нейросетевыми моделями. Виртуальные машины были сконфигурированы с учётом требований к оперативной памяти и числу ядер процессора. Дополнительным преимуществом Yandex.Cloud стало наличие образовательного гранта, позволившего использовать ресурсы без дополнительных затрат.

Для предобработки и анализа текстов применялись библиотеки NLTK и Natasha, обеспечивающие инструменты для токенизации, нормализации и лемматизации. Для тематической классификации текстов использовались алгоритмы из библиотеки scikit-learn:

1. k-NN (k ближайших соседей) — простой и интуитивно понятный метод, хорошо работающий на небольших выборках.
2. SGDClassifier — линейная модель, обучающаяся с помощью стохастического градиентного спуска, эффективно справляется с большими объёмами данных.

3. Random Forest — ансамблевый метод, объединяющий несколько решающих деревьев и устойчивый к переобучению.
4. Логистическая регрессия — один из базовых методов классификации, отличающийся высокой интерпретируемостью.

Для обучения и инференса генеративных моделей были задействованы возможности библиотеки Transformers, разработанной компанией HuggingFace. Также использовались средства визуализации для анализа результатов и построения графиков, позволяющих оценить поведение моделей на разных этапах обучения.

Третий раздел «Программная реализация методов суммаризации текстов» охватывает весь практический цикл от подготовки данных до анализа полученных результатов.

В качестве основного источника данных был выбран Mixed-Summarization-Dataset — совокупность нескольких русскоязычных датасетов: XLSum, Gazeta, WikiLingua, MLSUM, Reviews (ru), Curation-corpus (ru), Matreshka, DialogSum (ru), SAMSum (ru) [6]. Данные включают новостные тексты, диалоги и инструкции с краткими аннотациями.

На первом этапе выполнены загрузка и очистка данных. Для повышения точности обучения и последующей оценки качества суммаризации датасет был тематически сегментирован. Тематическое разбиение позволяет учитывать особенности разных типов текстов: например, юридический или медицинский стиль требует большей точности и устойчивости к искажению структуры, чем разговорный стиль.

Для классификации по темам были обучены модели на базе упомянутых ранее алгоритмов классификации. В качестве обучающей выборки использовались тексты новостных статей, полученные через API NewsAPI.

После подготовки и тематического разделения данных были дообучены две модели: T5 и BART. Обучение происходило на соответствующих подвыборках с использованием CPU, что повлияло на длительность обучения.

Завершающим этапом стала сравнительная оценка результатов. Модель BART продемонстрировала более высокие значения по всем ключевым метрикам: ROUGE, METEOR, BLEU и CHRF. Также качественный анализ выходных текстов показал, что генерации BART отличались большей связностью и смысловой точностью. Это объясняется большей архитектурной глубиной модели и числом параметров, что позволяет BART лучше справляться с задачами обобщения и генерации текста.

ЗАКЛЮЧЕНИЕ

В ходе работы был проведён анализ современных методов автоматической абстрактивной суммаризации текстов на русском языке. Для оценки качества использовались метрики ROUGE-n, ROUGE-L, METEOR, CHRF и BLEU, что позволило получить объективную и разностороннюю оценку результатов. В качестве основного источника данных использовался один русскоязычный датасет, который был разделён на тематические группы с помощью методов классификации для повышения качества обучения моделей. Были выбраны и дообучены две предобученные генеративные модели – BART и T5. В ходе работы было выявлено, что модель BART превосходит T5 как по значениям метрик качества суммаризации, так и по качеству сгенерированных текстов. Такой результат обусловлен большей архитектурной сложностью и числом параметров BART, что обеспечивает более эффективную обработку и обобщение исходной текстовой информации.

Основные источники информации:

1. Пересказ текстов в Яндекс Браузере [Электронный ресурс] URL: https://browser.yandex.ru/c/summarization (Дата обращения: 22.02.2025)

2. Summary Translation + Localization: Top Strategic Technology Trends for 2022: Generative AI [Электронный ресурс] URL: https://www.gartner.com/en/documents/4009656 (Дата обращения: 22.02.2025)

3. Towards Russian Summarization: can architecture solve data limitations problems [Электронный ресурс] URL: https://sberlabs.com/static/files/1003/RU/Russian_Summarization__Camera_ready_.pdf?ysclid=maashft5ge743462331 (Дата обращения: 22.02.2025)

4. Aman Kedia, Mayank Rasu. Hands-On Python Natural Language Processing – Birmingham, UK: 2020. – 60-68с.

5. Raffel, Colin, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Journal of Machine Learning (2020) – С. 1-67 [Электронный ресурс]: URL: https://arxiv.org/abs/1910.10683 (Дата обращения: 22.02.2025)

6. Hugging Face RussianNLP/Mixed-Summarizatio-Dataset [Электронный ресурс] URL: https://huggingface.co/datasets/RussianNLP/Mixed-Summarization-Dataset (Дата обращения: 18.04.2025)