Автоматическая суммаризация новостных статей для русского языка

Березовский Станислав
Студент группы DS-80

Содержание

Введение

1

Актуальность темы. NLP и ATS

- Ежедневный рост объема информации
- Важность обработки естественного языка (NLP) в современном мире
- Автоматическая суммаризация текста (ATS) как решение проблемы информационной перегрузки

Теоретические основы

2

Основные подходы к ATS

Extractive summarization

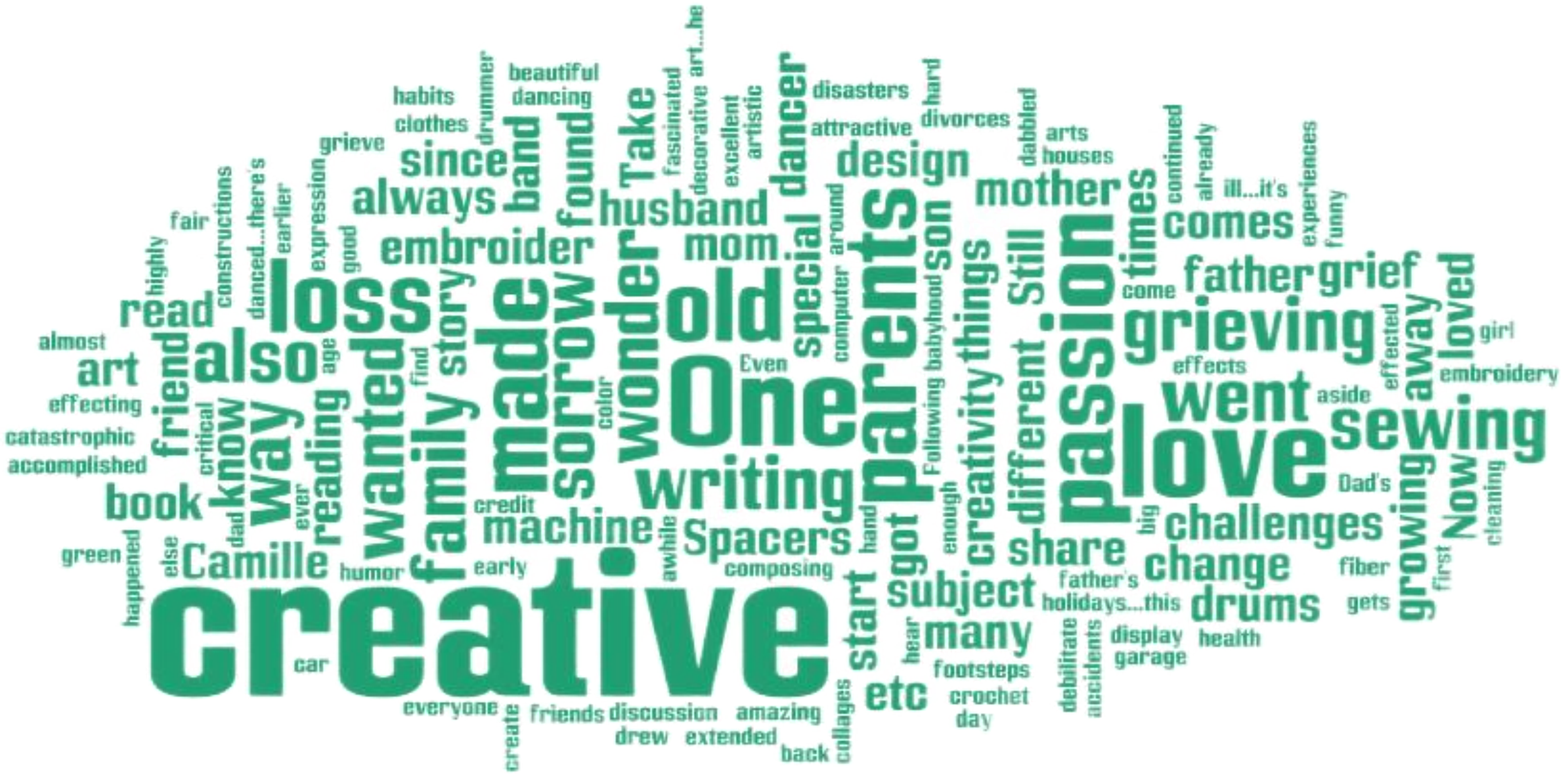
Заключается в выборе ключевых предложений или фраз из исходного текста, которые затем объединяются для создания суммарного текста. Экстрактивные методы просты в реализации и требуют меньше вычислительных ресурсов, однако их результат может быть менее естественным и связным.

Abstractive summarization

Предполагает создание нового текста, который передает основную идею исходного документа, используя перефразирование и генерацию новых предложений. Абстрактивная суммаризация более сложна в реализации, но позволяет получить более естественный и читаемый текст.

Основные методы предобработки текста

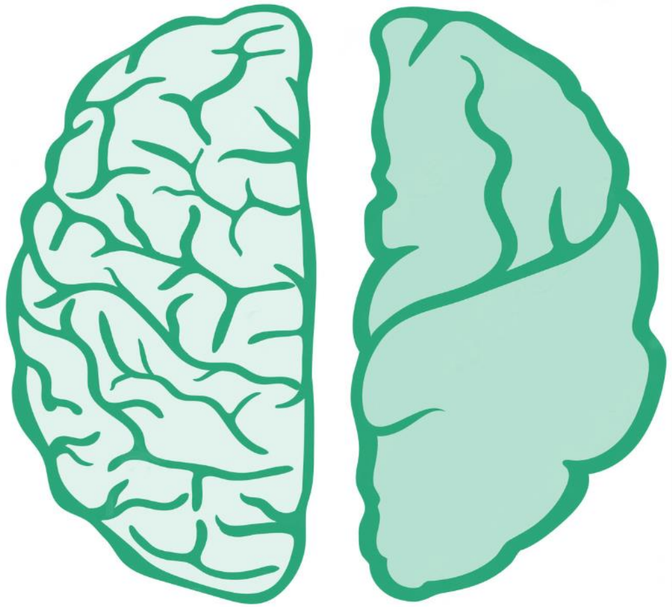
1. Приведение к нижнему регистру
2. Удаление знаков препинания
3. Удаление стоп-слов
4. Лемматизация
5. Стемминг
6. Токенизация



Основные методы ATS

Современные методы

Включают модели на основе глубокого обучения, такие как трансформеры. Они позволяют учитывать контекст и генерировать более осмысленные и естественные суммаризации. Методы, основанные на глубоком обучении, характеризуются тем, что они могут создавать краткие содержания, которые по смыслу близки к тем, что составлены человеком.



Традиционные методы

Включают частотный анализ и методы на основе графов. Частотный анализ основан на подсчете частоты появления слов в тексте и выборе наиболее частотных слов для создания резюме. Методы на основе графов используют графовые структуры для представления текста и выбора ключевых предложений.

Анализ и подготовка данных

3

Датасет "Gazeta"

Датасет "Gazeta" представляет собой коллекцию новостных статей с сайта www.gazeta.ru. Каждая запись включает URL новости, полный текст, заголовок, краткое резюме и дату публикации. Датасет разделён на тренировочный, тестовый и валидационный наборы, что позволяет корректно обучить и оценить модель.


```
{
  "url": "https://www.gazeta.ru/science/2020/02/14_a_12960289.shtml",
  "text": "Американское аэрокосмическое агентство NASA огласило
названия четырех космических миссий...",
  "title": "Венера, Ио или Тритон: куда полетит NASA",
  "summary": "В NASA назвали четыре миссии в дальний космос...",
  "date": "2020-02-14 16:39:11"
}
```

Статистический анализ данных

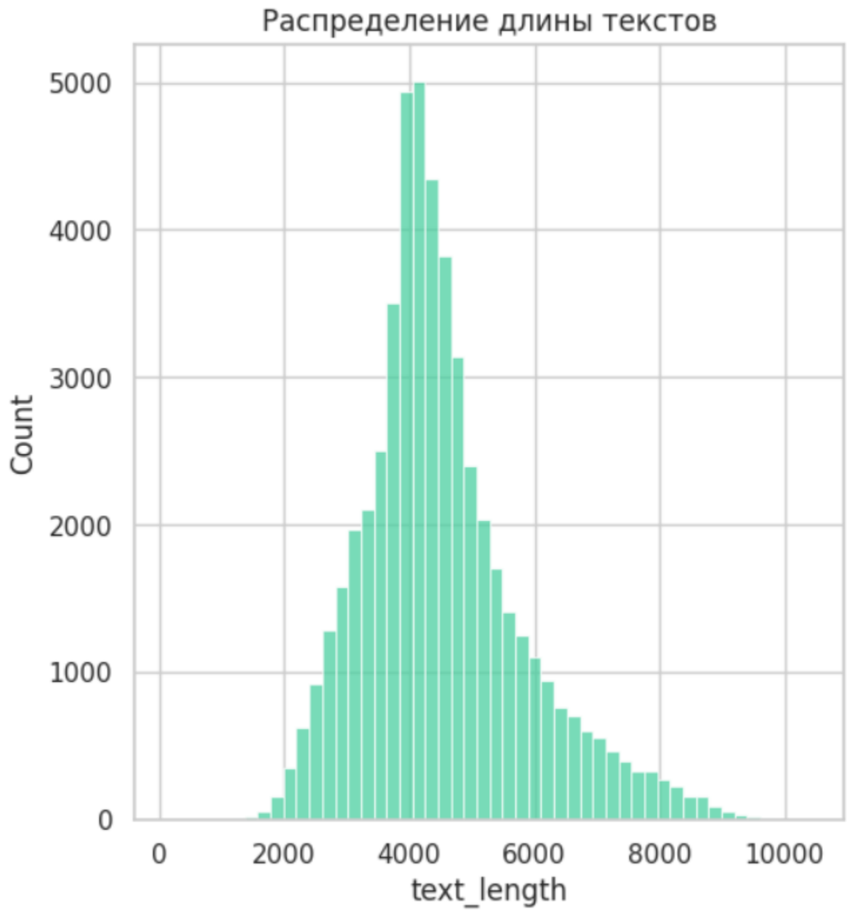
Статистика набора данных после перевода текстов в нижний регистр

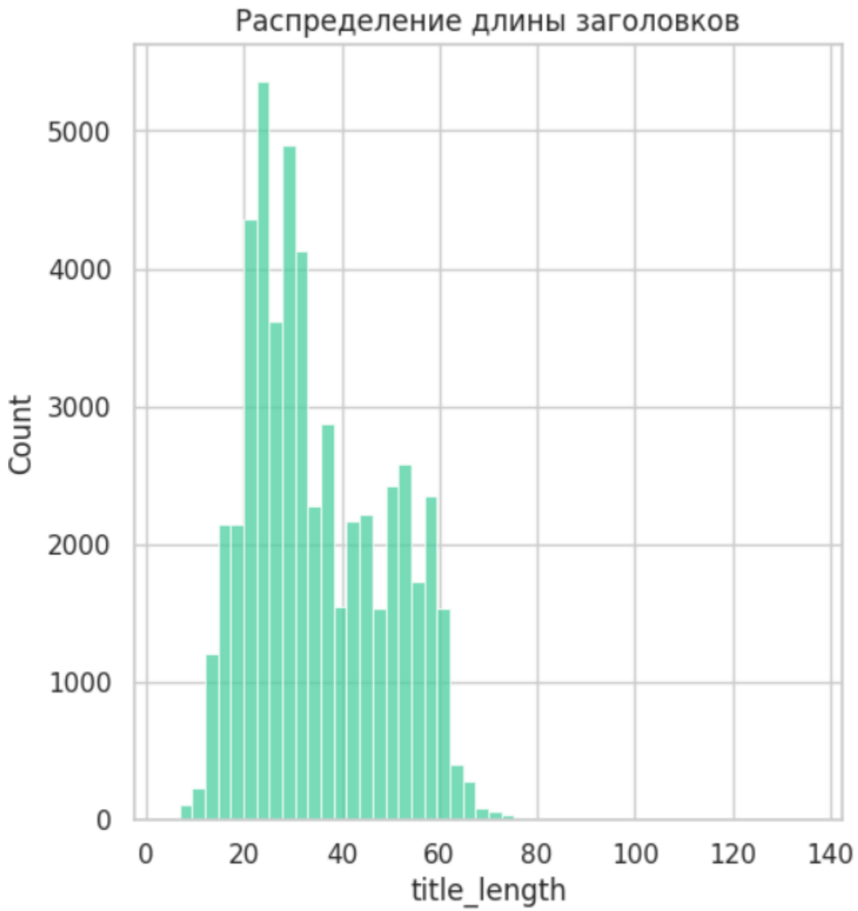
	Train		Validation		Test	
	Text	Summary	Text	Summary	Text	Summary
Pairs	52 400		5265		5770	
UW	611 829	148 073	167 612	42 104	175 369	44 169
UL	282 867	63 351	70 210	19 698	75 214	20 637
Common UL	60 992		19 138		20 098	
Min words	28	15	191	18	357	18
Max words	1500	85	1500	85	1498	85
Avg words	766.5	48.8	772.4	54.5	750.3	53.2
Avg UW	419.1	41.3	424.2	46.0	415.7	45.1
Avg UL	350.0	40.2	352.5	44.6	345.4	43.9

UW – Unique words, UL – Unique lemmas

Статистический анализ данных

Построим гистограммы распределения длины текстов и заголовков





Предварительная обработка данных

4

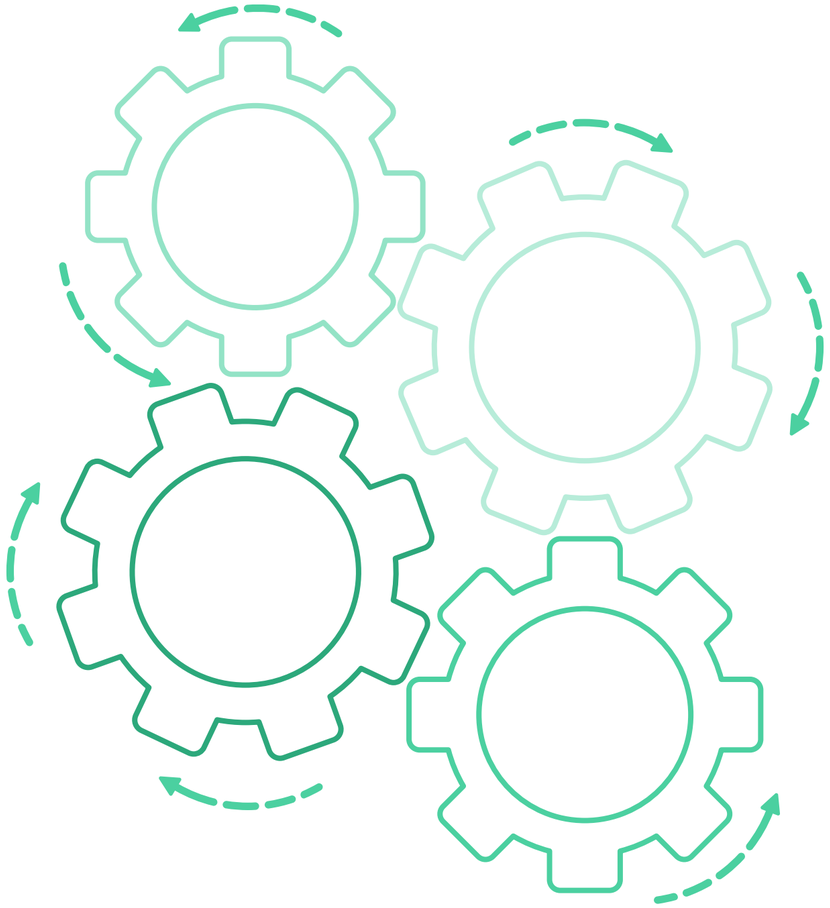
Подготовка текста

Удаление лишних символов

Удаление HTML-тегов, пунктуации и приведение текстов к нижнему регистру позволяет добиться унификации текста и упрощения его анализа.

Удаление стоп-слов

Стоп-слова — это слова, которые часто встречаются в тексте, но не несут значимой смысловой нагрузки и могут быть исключены из анализа текста. В основном, это предлоги, союзы, местоимения и другие служебные части речи.



Токенизация

Токенизация — это процесс разбиения текста на более мелкие единицы, называемые токенами. Этот процесс является фундаментальным шагом в NLP, так как он преобразует текст в формат, который нейронные сети могут анализировать и обучаться на нём.

Лемматизация

Лемматизация — процесс приведения словоформы к лемме — её нормальной (словарной) форме

Пример предобработанных данных

N°	Title	Cleaned title	Tokenized title	Title without stop words	Lematized title
1	Прогноз не успевает за оттоком	прогноз не успевает за оттоком	[прогноз, не, успевает, за, оттоком]	[прогноз, успевает, оттоком]	[прогноз, успевать, отток]
2	Google закончил поиск	google закончил поиск	[google, закончил, поиск]	[google, закончил, поиск]	[google, закончить, поиск]
3	«Фигуранты дела могут давить на свидетелей»	фигуранты дела могут давить на свидетелей	[фигуранты, дела, могут, давить, на, свидетелей]	[фигуранты, дела, могут, давить, свидетелей]	[фигурант, дело, мочь, давить, свидетель]
4	«С последних традиционно «отжимают» больше»	с последних традиционно отжимают больше	[с, последних, традиционно, отжимают, больше]	[последних, традиционно, отжимают]	[последний, традиционно, отжимать]
5	Третий «Голос» за Градского	третий голос за градского	[третий, голос, за, градского]	[третий, голос, градского]	[третий, голос, градский]

Метрики для оценки качества суммаризации

5

ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) - метрика, сравнивающая n-граммы, лексические цепочки и последовательности слов между сгенерированным и эталонным текстами.

$$\text{ROUGE - 1} = \frac{\sum_i Count_{match}\left(u_i\right)}{\sum_i Count_{ref}\left(u_i\right)}$$

$$\text{ROUGE - 2} = \frac{\sum_i Count_{match}\left(b_i\right)}{\sum_i Count_{ref}\left(b_i\right)}$$

$$\text{ROUGE - L} = \frac{LCS\left(S, G\right)}{\max\left(|S|, |G|\right)}$$

BLEU

BLEU (Bilingual Evaluation Understudy) - метрика для оценки качества машинного перевода, основанная на количестве совпадающих n-грамм в переведённом машиной и эталонном текстах.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$p_n = \frac{\sum_{i=1}^{N} \sum_{n-gram \in c_i} Count(n-gram)_{r_i}}{\sum_{i=1}^{N} \sum_{n-gram \in c_i} Count(n-gram)}$$

$$\text{BP} = \begin{cases} 1, \ if \ c > r \\ e^{(1-r/c)}, \ if \ c \leq r \end{cases}$$

BERTScore

BERT (Bidirectional Encoder Representations from Transformers) - метрика, использующая модель BERT при оценке схожести текстов, сравнивая их векторные представления для определения семантической близости.

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \underbrace{\max_{x_i \in x} \overbrace{x_i^T \hat{x}_j}^{\text{cosine similarity}}}_{\text{greedy matching}}$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \underbrace{\max_{\hat{x}_j \in \hat{x}} \overbrace{x_i^T \hat{x}_j}^{\text{cosine similarity}}}_{\text{greedy matching}}$$

$$F_{BERT} = 2 \cdot \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Разработка базовой модели

6

Baseline

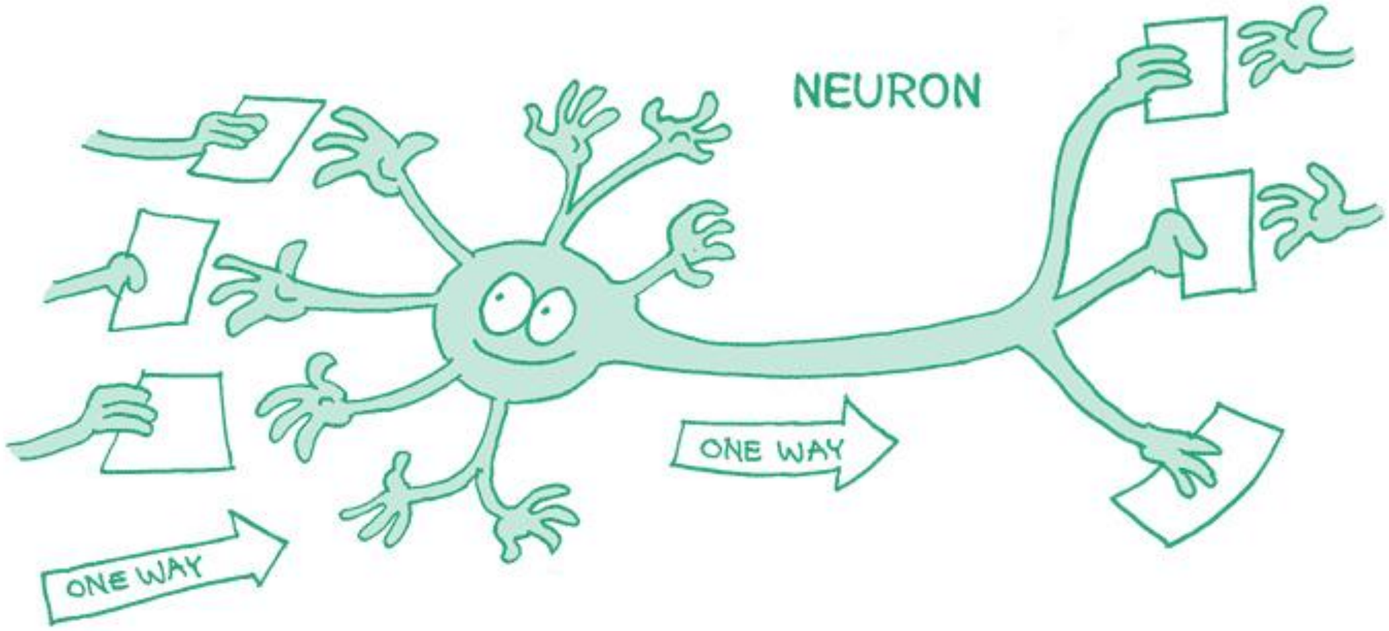
В качестве базовой модели использовалась модель для экстрактивной суммаризации, которая выбирает первое предложение статьи и использует его в качестве заголовка. Этот подход позволяет получить начальные результаты для сравнения с более сложными моделями.

Разработка продвинутой модели

7

Выбор модели для суммаризации

Transformer-based архитектуры показали превосходные результаты в задачах обработки естественного языка. Эти модели могут эффективно обрабатывать длинные тексты и сохранять контекст на больших расстояниях, что особенно важно для генерации заголовков. В данной работе была выбрана модель ruT5-base. Этот выбор обоснован её высокой производительностью и возможностью эффективной генерации текстов на русском языке.



Анализ результатов

8

ROUGE, BLEU и BERTScore

Видим, что удалось улучшить метрики BERTScore и BLEU, а вот значения метрики ROUGE ухудшились. Однако, уменьшение метрики ROUGE не означает снижение качества модели. Разница в метриках указывает на то, что модель ruT5-base лучше передаёт общий смысл и точность соответствия с точки зрения семантики, но менее точно совпадает с эталонными заголовками по конкретным словам и фразам.

Model	BERTScore	ROUGE-1-F	ROUGE-2-F	ROUGE-L-F	BLEU
baseline	67.557621	6.742888	0.742854	6.669108	1.791344
ruT5-base	70.542961	3.893125	0.280185	3.856153	2.761424

Заключение

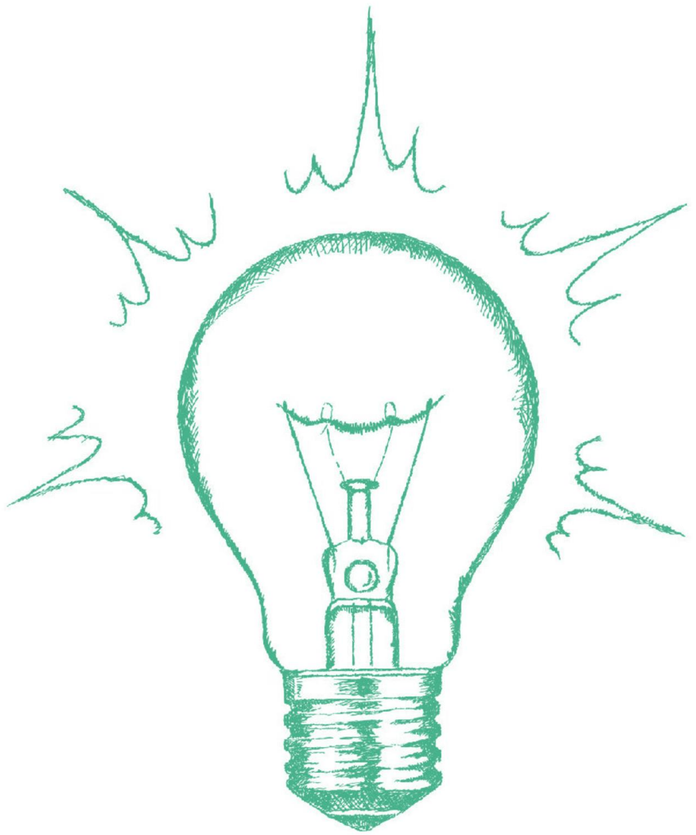
9

Резюме выполненной работы

В данной работе была разработана модель для автоматической генерации заголовков русскоязычных новостных статей на основе архитектуры трансформеров. Были рассмотрены различные подходы к суммаризации текста, проведён анализ данных и обучена модель ruT5-base. Качество полученных заголовков было оценено с использованием метрик ROUGE, BLEU и BERTScore. Получила подтверждение эффективность подходов на основе глубокого обучения для задач суммаризации. Модель показала неплохие результаты при генерации заголовков, однако некоторые ошибки указывают на необходимость дальнейшего улучшения и тонкой настройки модели.

Направления для будущих исследований

В будущем возможно улучшение модели за счёт использования более объёмных и разнообразных данных, а также применения новых архитектур и методов глубокого обучения для повышения качества генерируемых заголовков.



Спасибо за внимание!