Научная статья

УДК 81’322.2
DOI 10.25205/1818-7935-2025-23-1-80-92

Автоматическая саммаризация родительских чатов в WhatsApp

Кристина Александровна Дмитриева [1]
Марина Романовна Жолус [2]

Национальный исследовательский университет «Высшая школа экономики»
Санкт-Петербург, Россия

[1] kadmitrieva@hse.ru; https://orcid.org/0009-0001-9548-3273
[2] mrzholus@edu.hse.ru; https://orcid.org/0009-0005-4124-1956

Аннотация

Автоматическая саммаризация текста – одна из ключевых задач NLP, предполагающая создание краткой версии исходного текста. В современном мире, где объемы потребляемой человеком информации неустанно растут, задаче саммаризации уделяется все больше внимания. Автореферирование предполагает два основных подхода: экстрактивный и абстрактивный. Последний заключается в автоматическом создании саммари текста, в котором могут содержаться слова и предложения, не встречающиеся в источнике. Этот подход зачастую требует использования нейросетевых моделей, и для его реализации необходимы большие наборы специальным образом размеченных данных. Несмотря на значительные успехи в абстрактивной саммаризации публицистических и научных текстов, методы и датасеты, используемые для работы с монологическими документами, не всегда применимы для саммаризации диалогов. Кроме того, хотя создано достаточно много англоязычных датасетов для саммаризации текстов различных доменов, существующие наборы данных для автоматического аннотирования текстов на русском языке пока немногочисленны. Настоящая статья посвящена разработке и описанию русскоязычного диалогового датасета для саммаризации сообщений в родительских чатах и последующему обучению модели абстрактивной саммаризации для русского языка на авторском наборе диалоговых данных. В качестве материала выступил родительский чат с учителем в мессенджере WhatsApp. Процесс ручной разметки датасета включал в себя разбиение всех сообщений чата на отдельные диалоги, создание саммари и присвоение тематических меток для каждого разговора. В результате был создан датасет, содержащий 616 диалогов, в общей сложности состоящих из 3380 сообщений. Для файн-тьюнинга были выбраны модели-трансформеры ruT5, mT5 и RuGPT (ruT5 и RuGPT были предварительно обучены на русскоязычном датасете для автоматической саммаризации новостей), а для оценки их качества – метрики ROUGE-1, ROUGE-2, ROUGE-L, BLEU и BERTScore. В результате модели ruT5, дообученной на авторском датасете, удалось превзойти бейзлайн по всем пяти метрикам.

Ключевые слова

автоматическая саммаризация текста, диалоговая саммаризация, машинное обучение, трансформеры, обработка естественного языка.

Финансирование

Исследование подготовлено по материалам проекта «Текст как Big Data: методы и модели работы с большими текстовыми данными», выполняемого в рамках Программы фундаментальных исследований НИУ ВШЭ в 2024 году.

Для цитирования

Дмитриева К. А., Жолус М. Р. Автоматическая саммаризация родительских чатов в WhatsApp // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2025. Т. 23, № 1. С. 80–92. DOI 10.25205/1818-7935-2025-23-1-80-92

Automatic Summarization of Parental Chats on WhatsApp

Kristina A. Dmitrieva[1], Marina R. Zholus[2]

HSE University, Saint Petersburg,
Russian Federation

[1] kadmitrieva@hse.ru; https://orcid.org/0009-0001-9548-3273
[2] mrzholus@edu.hse.ru; https://orcid.org/0009-0005-4124-1956

Abstract

Automatic text summarization is one of the main tasks of natural language processing (NLP), which consists in creating a shorter version of the source text. In today's world the amount of information consumed by people is constantly increasing, therefore more and more emphasis is being placed on the task of summarization. There are two main approaches to automatic text summarization: extractive and abstractive ones. The latter involves automatic creation of a summary text that may contain words and phrases not present in the source. This approach usually requires the usage of AI models, which creates a demand for large datasets labeled in a certain way. Despite significant advances in summarization of scientific and news articles, the methods and datasets applied to monologue documents are not always suitable for dialogue summarization. Besides, although there exists a considerable number of English-language summarization datasets, the number of those available in Russian is not yet sufficient. The paper is devoted to the labeling and description of a Russian-language dataset for group chat messages summarization and fine-tuning models for the task of abstractive summarization for the Russian language on a custom dialogue dataset. A parental chat with a teacher in WhatsApp was used as material for the dataset. The process of manually labeling the dataset consisted in dividing the entire group chat into separate dialogues, writing a summary, and adding topic labels for each of them. As a result, a dataset has been created, which includes 616 dialogues with a total of 3380 messages. The ruT5, mT5 and RuGPT models were selected for fine-tuning, the ruT5 and RuGPT models were pre-trained on a Russian-language dataset for automatic news summarization. The ROUGE–1, ROUGE-2, ROUGE-L, BLEU and BERTScore metrics were used to evaluate the quality of the models. Subsequently, the ruT5 model, fine-tuned on the custom dataset, turned out to outperform the baseline model in all the five metrics.

Keywords

Automatic text summarization, dialogue summarization, machine learning, transformers, dataset, NLP

Funding

The article is based on the materials of the project "Text as Big Data: methods and models of working with large text data", carried out within the framework of the HSE Fundamental Research Program in 2024

For citation

Dmitrieva K. A., Zholus M. R. Automatic Summarization of Parental Chats on WhatsApp. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2025, vol. 23, no. 1, pp. 80–92. DOI 10.25205/1818-7935-2025-23-1-80-92

Введение

Саммаризация текста – это задача автоматической генерации краткого и емкого резюме, отражающего основные идеи текста-источника [Liu et al., 2018]. Первые попытки создания сокращенных версий текстовых документов с помощью компьютерных программ были предприняты еще в 50-х годах прошлого века [Luhn, 1958]. С тех пор объемы и скорость накопления информации человечеством значительно увеличились, что сделало автоматическую саммаризацию текстов одной из важнейших задач NLP (Natural Language Processing) [Narayan et al., 2018: 1797]. Появились новые методы решения этой задачи: наряду с экстрактивным подходом, предполагающим создание аннотаций из ключевых предложений исходного текста [Liu, 2018], получила широкое распространение абстрактивная саммаризация, допускающая использование в саммари слов и предложений, не встречающихся в источнике [Gliwa et al., 2019: 70]. Поскольку абстрактивный подход предполагает генерацию текста и работу с последовательностями, для его реализации обычно применяют рекуррентные нейронные сети или модели-трансформеры [Jin et al., 2024].

Исследователи добились значительных успехов в автоматической генерации саммари для монологических текстов разных доменов: новостей [Hermann et al., 2015], правовых документов [Shukla et al., 2022], научных публикаций [An et al., 2021: 12498], медицинских вопросов [Ghosh et al., 2024]. Однако методы, эффективные для резюмирования монологических текстов, хуже справляются с саммаризацией диалогической речи [Rameshkumar, Bailey, 2020: 5123]. Одна из причин этого явления – сравнительно небольшое количество существующих датасетов для абстрактивной саммаризации диалогов, в особенности на русском языке. Хотя работы по созданию диалоговых датасетов для саммаризации текстов различной тематики ведутся с начала XXI века, их общее количество и объем размеченных данных значительно уступают датасетам новостных статей и научных публикаций [Gliwa et al., 2019: 70].

В настоящей статье описан процесс создания русскоязычного датасета для абстрактивной диалоговой саммаризации сообщений из родительских чатов, а также последующее дообучение саммаризационных моделей на получившемся наборе размеченных данных. Выбранный текстовый домен представляется нам интересным не только ввиду диалогической природы, но и из-за актуальной тематики. Родительские чаты в социальных сетях и мессенджерах стали настолько заметным явлением в современном российском обществе, что в 2021 году ВЦИОМ (Всероссийский центр изучения общественного мнения) провел исследование, посвященное этому феномену. Сегодня родители тратят большое количество времени на чтение переписки в школьных чатах, испытывая при этом трудности в преодолении конфликтных ситуаций [Bylieva, Lobatyuk, 2023]. Созданный датасет и обученные модели для саммаризации родительских чатов призваны упростить и ускорить извлечение полезной информации при взаимодействии родителей и учителей в социальных сетях и мессенджерах.

Обзор литературы

В 1958 году Лун [Luhn, 1958] одним из первых реализовал алгоритм саммаризации на материале научных публикаций и назвал новую задачу «автоматическим созданием аннотаций», основной целью исследователя было создание модели, способной находить ключевые предложения статей и формировать из них авторефераты. Позже подобный метод, заключающийся в поиске главных предложений, содержащих основные сведения из исходного текста, получил название «экстрактивный» (extractive). Основное преимущество экстрактивной саммаризации заключается в том, что для ее реализации можно использовать так называемые алгоритмы машинного обучения без учителя, а значит, сэкономить значительные ресурсы на разметке данных [Moratanch, Gopalan, 2017]. С распространением нейронных сетей и новых методов саммаризации этот подход стал менее популярен, однако до сих пор применяется в некоторых исследованиях [Chowdhury et al., 2024].

Абстрактивная саммаризация – сравнительно новый подход к задаче автоматического создания аннотаций, реализованный во многих современных исследованиях [Zhang et al., 2021; Lyu et al., 2024]. Он предполагает генерацию текстов саммари с использованием слов и предложений, которые не содержатся в исходном тексте. Хотя этот метод позволяет создавать более естественные и емкие резюме, для обучения нейросетевых моделей для абстрактивной саммаризации необходимы большие объемы качественно размеченных данных (датасеты, содержащие исходный текст и эталонное саммари), что усложняет применение абстрактивного подхода на практике [Jin et al., 2024].

С середины прошлого века не только появились новые методы автоматической саммаризации текстов, но и расширилась сфера ее применения. Так, помимо научных публикаций, саммаризацию применяют к юридическим документам [Shukla et al., 2022], новостным статьям [Hermann et al., 2015], диалогам [Feng et al., 2022], медицинским текстам [Ghosh et al., 2024], постам из блогов и соцсетей [Dutta et al., 2018] и текстам других жанров и тематик.

На сегодняшний день существует достаточное количество наборов данных для решения задачи саммаризации монологических текстов разных доменов на английском языке. Один из крупнейших таких датасетов – Gigaword [Napoles et al., 2012], состоящий из 9 876 086 новостных статей с заголовками в качестве саммари. Несмотря на очевидное преимущество в количестве размеченных текстов, датасет подвергается критике за использование заголовков в качестве саммари [Jin et al., 2024]. Еще один не менее популярный датасет для саммаризации новостей – корпус статей, собранных с официальных сайтов CNN и DailyMail, с их абстрактивными саммари [Hermann et al., 2015]. Всего этот набор данных содержит 312 085 пар текст-саммари, и ввиду своего объема часто применяется для решения задач саммаризации текстов других доменов: в частности, создатели диалоговых датасетов [Gliwa et al., 2019; Chen et al., 2021] используют модели, обученные на корпусе CNN/DailyMail, как бейзлайн при оценке своих данных или объединяют свои примеры с тренировочной частью новостного датасета при обучении нейронных сетей для достижения лучшего качества саммаризации. Среди крупных датасетов с текстами других жанров можно выделить корпус научных публикаций, собранных с порталов arXiv.org (215 000 статей) и PubMed.com (133 000 статей), с авторскими аннотациями в качестве эталонных саммари [Cohan et al., 2018: 618].

Русскоязычные наборы данных для решения задачи саммаризации менее разнообразны и в основном представлены текстами газетных статей с заголовками или специально созданными аннотациями в качестве саммари (например, датасет Gazeta [Gusev, 2020]). Кроме того, существуют крупные мультиязычные наборы данных для задач саммаризации, в которых встречаются в том числе тексты на русском языке, например, XL-Sum [Hasan et al., 2021].

Датасеты для диалоговой саммаризации

Диалоговая саммаризация как отдельное направление прикладных исследований требует других наборов данных для обучения моделей. Хотя в последние годы появилось достаточно много новых диалоговых датасетов, количество и общий объем наборов данных для саммаризации монологических текстов значительно превышает те же показатели датасетов для диалоговой саммаризации. При этом многие из существующих диалоговых датасетов посвящены решению задачи саммаризации текстов конкретного домена, например, созданию медицинских выписок (MTS-Dialiog [Ben Abacha et al., 2023]) или кратких выжимок диалогов с клиентами (DiDi [Liu et al., 2019]). В табл. 1 представлены обобщенные сведения об основных датасетах для диалоговой саммаризации, отсортированные по количеству диалогов.

Таблица 1

Сравнение датасетов для диалоговой саммаризации

Table 1

Comparison of dialogue summarization datasets

Название Name	Год Year	Количество диалогов Number of dialogues	Домен Domain
1	2	3	4
MediaSum [Zhu et al., 2021]	2021	463 596	Интервью
DiDi [Liu et al., 2019]	2019 (нет в публичном доступе)	328 880	Обращения в службу поддержки
SumTitles [Malykh et al., 2020]	2020	21 479	Субтитры к фильмам

Окончание табл. 1

1	2	3	4
SAMSum [Gliwa et al., 2019]	2019	16 369	Переписка в чатах
DialogueSum [Chen et al., 2021]	2021	13 460	Повседневная разговорная речь
EmailSum [Zhang et al., 2021]	2021	2 549	Электронные сообщения
MTS-DIALOGUE [Ben Abacha et al., 2023]	2023	1 701	Диалоги врачей с пациентами
TWEETSUM [Feigenblat et al., 2021]	2021	1 100	Обращения в службу поддержки
QMSum [Zhong et al., 2021]	2021	232	Совещания
CRD3 [Rameshkumar, Bailey, 2020]	2020	159	Словесные ролевые игры
AMI [Carletta et al., 2006]	2005	137	Совещания
ELITR [Nedoluzhko et al., 2022]	2022	120	Совещания
ISCI [Janin et al., 2003]	2003	75	Совещания

На основе данных из табл. 1 можно сделать вывод, что многие англоязычные датасеты для диалоговой саммаризации призваны решить задачу автоматического создания протоколов совещаний и конференций. Такие наборы данных в основном содержат небольшое количество текстов, однако средняя длина диалога в датасетах для создания протоколов совещаний значительно больше, чем в диалоговых корпусах других доменов [Zhu et al., 2021: 5930].

Единственный диалоговый датасет не на английском языке, представленный в таблице, – DiDi [Liu et al., 2019]. Кроме языка, он также примечателен своим объемом: 328 880 диалогов. Тексты представляют собой разговоры клиентов платформы транспортных услуг DiDi со службой поддержки, а эталонные аннотации написаны сотрудниками службы, участвовавшими в разговоре. Поскольку данные составляют коммерческую тайну, датасет не был опубликован.

Множество новых экспериментов в области диалоговой саммаризации появилось после публикации SAMSum [Gliwa et al., 2019] – первого датасета для задачи абстрактивной саммаризации онлайн-чатов. Он находится в открытом доступе и содержит эталонные аннотации, составленные лингвистами специально для задачи саммаризации. Интересно также, что сами диалоги были созданы искусственно специально для датасета и не являются реальными примерами сообщений из чатов. SAMSum критикуют за сравнительно короткие диалоги (в среднем 94 токена) и ограниченное количество затронутых в переписках тем [Chen et al., 2021: 5063].

Еще один популярный датасет для диалоговой саммаризации – DialogSum [Chen et al., 2021]. Он состоит из повседневных бесед на различные тематики, собранных с сайтов для изучения английского языка и аннотированных вручную. Создатели отмечают, что их диалоги длиннее и тематически разнообразнее [Chen et al., 2021: 5063], чем в SAMSum, однако значения метрик качества моделей, обученных на DialogSum, стабильно хуже, чем показатели тех же моделей, обученных на SAMSum [Chen et al., 2021: 5066].

Такие образом, за последние 20 лет появилось достаточно много англоязычных диалоговых датасетов, был собран крупный корпус диалогов на китайском языке. Однако наборы размеченных данных для диалоговой саммаризации на русском языке пока немногочисленны и в основном представлены непубличными датасетами для создания протоколов совещаний

и машинными и полумашинными переводами англоязычных датасетов. Так, например, существуют русскоязычные версии корпусов SAMSum и DialogSum, но автоматически переведенные на русский язык саммари содержат ошибки и неточности в формулировках. Собранный нами датасет представляет набор размеченных данных для русскоязычной диалоговой саммаризации в домене родительских онлайн чатов.

Методы и материал исследования

В качестве основы для датасета выступил родительский чат с учителем в мессенджере WhatsApp. Чат был экспортирован в файл формата plain text при помощи встроенной функции в приложении, а далее преобразован нами в файл json-формата, в котором номера телефонов и имена отправителей сообщений были заменены на условные обозначения формата «Родитель_[индекс]» (например, «Родитель_9») и «Классный_руководитель», а также сохранены данные о дате и времени отправления сообщения.

Разметка данных была осуществлена вручную авторами работы. Для этого с опорой на существующие исследования [Gliwa et al., 2019; Chen et al., 2021] был разработан ряд правил, необходимых для унификации разметки и последующего качественного обучения модели:

- оригинальный датасет должен быть очищен от системных сообщений, таких как уведомления об удалении сообщений, добавлении участников в чат или отправке медиафайлов;
- членение диалога решено совершать с опорой на даты отправки сообщений таким образом, при котором каждая тема внутри отрезка диалога должна быть логически завершена, даже если параллельно обсуждается вторая: таким образом, один выделенный диалог может содержать несколько тем, обсуждаемых параллельно разными участниками чата;
- размеченные данные должны содержать поля «id», «text» с оригинальным отрывком разговора, состоящим из одного или нескольких сообщений, «summary» с кратким резюме разговора, «topics» с тематическими метками разговора и «type» – поле, в которое записывалось значение «монолог» или «диалог», исходя из количества человек, чьи реплики вошли в тот или иной отрезок переписки;
- аннотации должны быть написаны кратко, в третьем лице и в настоящем времени, без графических и текстовых эмодзи и в стиле, приближенном к официально-деловому;
- для каждого диалога разметчик составляет один вариант саммари, после чего каждое написанное резюме проходит процесс утверждения и редактуры (при необходимости) вторым аннотатором;
- допускается присвоение одному разговору нескольких тематических меток, длина каждой из которых не должна была превышать трех токенов;
- тематические метки присваиваются вручную разметчиком на основе анализа содержания исходного диалога, в дальнейшем метки от разных разметчиков проходят процесс унификации (темы, близкие по смыслу, но по-разному названные разметчиками, приводятся к одному тэгу).

После проведения разметки мы посчитали показатель BERTScore [Zhang et al., 2020] для исходных диалогов и саммари, чтобы подтвердить соответствие аннотации содержанию исходного текста с помощью формальных метрик. Получившееся значение (0,751) и экспертная оценка качества саммари со стороны второго разметчика дали нам основание для использования датасета при дообучении языковых моделей для задачи саммаризации диалогов на русском языке.

Размеченный набор данных мы разделили случайным образом на три выборки:

- обучающую: 70 %, 432 диалога;

- валидационную (использовавшуюся для подбора гиперпараметров модели): 15 %, 92 диалога;
- тестовую: 15 %, 92 диалога.

Далее на авторском датасете нами был проведен файн-тьюнинг моделей-трансформеров (RuGPT-3, ruT5 и mT5) через библиотеку transformers от Hugging Face, а также последующая валидация моделей с помощью метрик ROUGE [Lin, 2004], BLEU [Papineni et al. 2002] и BERTScore [Zhang et al., 2020] и соответствующих python-библиотек.

Результаты и дискуссия

Описание разработанного датасета

В результате проведенной работы был получен размеченный диалоговый датасет для задачи саммаризации, основные параметры которого представлены в табл. 2.

Таблица 2

Ключевые параметры разработанного датасета

Table 2

The dataset's key characteristics

Количество сообщений	3 380	
Количество разговоров	616	
Сравнение диалогов и саммари	Исходный диалог	Саммари
Уникальные слова	20 638	3 651
Уникальные леммы	7 847	1 982
Минимальное количество токенов в диалоге	5 (M30)	3 (M19)
Максимальное количество слов в диалоге	3 937 (M158)	160 (K169)
Среднее количество слов в диалоге	145,16	23,62
Медианное количество слов в диалоге	84,0	18,0

На рисунке представлен фрагмент размеченного набора данных.


```
{
      "id": "K129",
      "text": "Классный_руководитель:
https://forms.office.com/Pages/ResponsePage.aspx?id=br6Yju\nКлассный_руков
одитель: Уважаемые родители! По ссылке выше ⬆️ можно предварительно
записаться в объединения платных образовательных услуг\nРодитель_33:
Добрый день! Расписание уроков на 1 сент уже есть?\nКлассный_руководитель:
Добрый день.Нет расписания\nРодитель_23: Для записи в доп.кружки хотелось
бы знать время их проведения. Или потом можно будет отказаться, если по
времени не стыковка?\nКлассный_руководитель: 1 сентября в 8-15 ребят
встречаю около входа 2 (боковая калитка),в 8-45 первый урок-классный
час,дальше уроки по расписанию (смотрим в дневнике,как
узнаю-сообщу)\nКлассный_руководитель: Без основного расписания трудно
определить дни для доп.кружков,обществознание(Жарикова) в среду.Всё можем
сказать предварительно\n",
      "summary": "Классный руководитель приглашает записываться на платные
кружки, но Родитель_33 перед этим хочется ознакомиться с расписанием,
которого еще нет. Пока известно, что 1 сентября в 8:15 учеников ждут у
входа №2, а в 8:45 будет классный час.",
      "topics": ["расписание", "дополнительные занятия", "классный час"],
      "type": "диалог"
}
```


Фрагмент размеченного датасета
Sample of the labeled dataset

Помимо эталонных саммари, датасет содержит тематические метки, присвоенные авторами исследования каждому диалогу. Предполагается, что данные метки впоследствии могут быть применены для классификации диалогов по тематике. Кроме того, указание типа разговора у каждого элемента датасета позволяет как при необходимости отфильтровать данные таким образом, чтобы в них обязательно участвовало более одного участника, так и использовать датасет целиком, включая разговоры, состоящие из высказываний одного участника, что актуально для онлайн-чатов.

Мы провели также дополнительное исследование связи тем диалогов с качеством работы модели саммаризации. Для этого из тестовой выборки были отобраны диалоги по трем темам: «здоровье», «ссылка» и «домашнее задание». Данные темы были выбраны, поскольку в тестовую выборку попало примерно одинаковое и достаточное количество диалогов (10–11), относящихся к этим тематикам. Далее с помощью одной из наших моделей были сгенерированы саммари для этих диалогов, для каждой тематической группы посчитаны метрики качества саммаризации. Статистически значимых различий в метриках выявлено не было, при экспертном анализе саммари также не удалось выявить какие-либо характерные особенности реферирования диалогов разных тематик. Возможно, это связано с недостаточным количеством данных, и при расширении датасета мы бы наблюдали интересные закономерности, однако на данном этапе полученные результаты не демонстрируют зависимости качества саммаризации от характера темы саммаризируемого диалога.

Дообучение моделей саммаризации

После создания датасета мы приступили к файн-тьюнингу саммаризационных моделей RuGPT-3 Medium, ruT5 и mT5. Они были выбраны нами для экспериментов, поскольку работают с текстами на русском языке, часто используются в современных исследованиях в области автоматической саммаризации, а также не требуют слишком больших вычислительных ресурсов для файн-тьюнинга. Причем первые две модели дообучены для задачи саммаризации русскоязычных новостей на датасете Gazeta [Gusev, 2020] и содержат только русские и английские эмбеддинги, а третья была взята для сравнения, так как является мультиязычной и не была дообучена на русском новостном корпусе, что, с одной стороны, усложняет работу с ней, а с другой – исключает возможность получения некачественных резюме из-за ориентированности на новостной формат текстов. Все модели относятся к трансформерам – архитектуре, впервые представленной в 2017 году исследовательским коллективом компании Google [Vaswani et al., 2017].

Для оценки качества работы моделей было решено использовать метрики ROUGE [Lin, 2004], BLEU [Papineni et al. 2002] и BERTScore [Zhang et al., 2020]. В качества бейзлайна была выбрана упомянутая выше ruT5 как модель, показавшая в исследовании [Gusev, 2020] лучший результат для задачи саммаризации русскоязычных текстов по сравнению с RuGPT-3 Medium и mBART.

В табл. 3 представлено сравнение метрик бейзлайн-модели, полученных на диалоговом датасете, с результатами дообученных нами моделей.

Как видно по табл. 3, лучший результат показывает модель ruT5, предварительно обученная на задачу саммаризации на новостном датасете и являющаяся полноценным трансформером – ее показатели по метрикам пакета ROUGE значительно превышают показатели других двух моделей, также она обладает наивысшей оценкой по BERTScore. RuT5 уступает лишь разделяющей с ней одну архитектуру модели mT5 по показателю BLEU. Кроме того, ruT5 превзошла бейзлайн-модель по всем пяти метрикам.

Тем не менее при оценке качества абстрактивной саммаризации необходимо также вручную смотреть на сгенерированные моделью краткие содержания. В ходе экспериментов с ар-

Таблица 3

Оценка моделей

Table 3

Model evaluation

Архитектура	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU**	**BERTScore (F1)**
Baseline	0,293	0,132	0,286	0,064	0,696
RuGPT-3 Medium	0,138	0,038	0,124	0,019	0,671
mT5	0,344	0,161	0,344	**0,305**	0,772
ruT5	**0,378**	**0,172**	**0,373**	0,147	**0,791**

Таблица 4

Сравнение сгенерированных саммари с эталонным

Table 4

Comparison of generated and reference summaries

Модель	**Саммари**
Эталон	Классный руководитель сообщает о готовности фотографий класса, предлагает выбрать понравившиеся и просит кого-либо из родителей заключить договор и оплатить фотографии по реквизитам за весь класс. Посмотреть фото вживую можно 1 июня
Ru-GPT	Родители, чьи фотографии получат в школе, смогут посмотреть их в реальном времени на сайте школы
mT5	Классный_руководитель сообщает о получении фотографий 1 июня в школу
ruT5	Классный_руководитель сообщает о получении и оплате фотографий

хитектурами RuGPT-3 и T5 были получены как удачные с субъективной точки зрения саммари, так и не слишком. В табл. 4 представлены краткие содержания, сгенерированные моделями для одного из диалогов после файн-тьюнинга на датасете родительских чатов.

Представленные примеры отражают недостатки, характерные для той или иной модели. Так, GPT склонна галлюцинировать, т. е. искажать информацию из источника и допускать фактические ошибки. Например, в эталонном саммари (см. табл. 4) сказано, что фото можно будет посмотреть «вживую», однако в реферате, полученном с помощью Ru-GPT, предлагается ознакомиться с фотографиями на сайте школы. Обе модели серии T5 не допускают фактических ошибок, однако создают слишком короткие саммари, из-за чего упускают важную информацию: mT5 ничего не сообщает об оплате фотографий, ruT5 – о времени ознакомления. Подобного рода недостатки наблюдаются и в других сгенерированных на тестовой части датасета саммари, однако в целом можно сказать, что модели достаточно неплохо справляются с задачей реферирования текстовых диалогов.

Заключение

Таким образом, в данной статье были рассмотрены актуальные проблемы в области задачи автоматической саммаризации текста, представлен новый датасет, разработанный с целью закрыть пробелы в сфере диалоговой саммаризации, а также приведены метрики качества саммаризационных моделей, дообученных на собранном наборе данных.

Согласно метрикам и экспертному анализу полученных саммари, из трех моделей лучше всего себя показала ruT5, предварительно также обученная на наборе данных Gazeta для зада-

чи саммаризации новостных статей на русском языке. Ей удалось превзойти бейзлайн по всем пяти метрикам.

Результаты исследования демонстрируют, что саммаризация в целом и диалоговая саммаризация в частности являются интересными, актуальными и перспективными направлениями в современных исследованиях и разработке, однако в русскоязычном научном пространстве интерес к этой теме в основном связан с закрытыми коммерческими разработками, а существующие датасеты для автоматического аннотирования текстов на русском языке пока немногочисленны.

Дообученные модели, представленные в статье, могут быть использованы в прикладных проектах, таких как разработка приложения или чат-бота для саммаризации диалогов в WhatsApp. Кроме того, планируется расширение полученного диалогового датасета путем добавления сообщений из чатов другой тематики, например, рабочих, студенческих или дружеских.

Список литературы / References

An C., Zhong M., Chen Y., Wang D., Qiu X., Huang X. Enhancing Scientific Papers Summarization with Citation Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35(14), pp. 12498–12506. https://doi.org/10.1609/aaai.v35i14.17482

Ben Abacha A., Yim W., Fan Y., Lin T. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2291–2302. Available at: https://aclanthology.org/2023.eacl-main.168.pdf (accessed: June 23, 2024).

Budzianowski P., Wen T., Tseng B.H., Casanueva I., Ultes S., Ramadan O., et al. MultiWOZ – A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5016–5026. https://doi.org/10.18653/v1/d18-1547

Bylieva D., Lobatyuk V., Novikov M. Parent Chats in Education System: During and after the Pandemic Outbreak. *Education Sciences*, 2023, vol. 13(8), pp. 778–794. https://doi.org/10.3390/educsci13080778

Carletta J., Ashby S., Bourban S., Flynn M., Guillemot M., Hain T., et al. The AMI Meeting Corpus: A Pre-announcement. *Lecture Notes in Computer Science*, 2006, pp. 28–39. https://doi.org/10.1007/11677482_3

Chen Y., Liu Y., Chen L., Zhang Y. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. Findings of the Association for Computational *Linguistics: ACL-IJCNLP* 2021, 2021, pp. 5062–5074. Available at: https://aclanthology.org/2021.findings-acl.449.pdf (accessed: June 24, 2024).

Chowdhury S.B.R., Monath N., Dubey A., Zaheer M., McCallum A., Ahmed A., Chaturvedi S. Incremental Extractive Opinion Summarization Using Cover Trees. arXiv (Cornell University). 2024. Available at: https://arxiv.org/abs/2401.08047 (accessed: June 24, 2024).

Cohan A., Dernoncourt F., Kim D. S., Bui T., Kim S., Chang W., Goharian N. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018, vol. 2, pp. 615–621. https://doi.org/10.18653/v1/n18-2097

Dutta S., Chandra V., Mehra K., Ghatak S., Das A. K., Ghosh S. Summarizing Microblogs during Emergency Events: A Comparison of Extractive Summarization Algorithms. *International Conference on Emerging Technologies in Data Mining and Information Security*, 2018. Available at: https://www.researchgate.net/publication/325593717_Summarizing_Microblogs_during_Emer-

gency_Events_A_Comparison_of_Extractive_Summarization_Algorithms (accessed: June 25, 2024).

Feigenblat G., Gunasekara R. C., Sznajder B., Joshi S., Konopnicki D., Aharonov R. TWEETSUMM A Dialog Summarization Dataset for Customer Service. *Findings of the Association for Computational Linguistics: EMNLP* 2021, 2021, pp. 245–260. https://doi.org/10.18653/v1/2021.findings-emnlp.24

Feng X., Feng X., Qin B. A Survey on Dialogue Summarization: Recent Advances and New Frontiers. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, pp. 5453–5460. https://doi.org/10.24963/ijcai.2022/764

Ghosh A., Acharya A., Jha P., Gaudgaul A., Majumdar R., Saha S., Chadha A., Jain R., Sinha S., Agarwal S. MedSUMM: A Multimodal Approach to Summarizing Code-Mixed Hindi-English clinical queries. arXiv (Cornell University). 2024. Available at: https://arxiv.org/abs/2401.01596 (accessed: June 24, 2024).

Gliwa B., Mochol I., Biesek M., Wawer A. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 2019, pp. 70–79. https://doi.org/10.18653/v1/d19-5409

Gusev I. Dataset for Automatic Summarization of Russian News. In: *Communications in computer and information science*, 2020, pp. 122–134. https://doi.org/10.1007/978-3-030-59082-6_9

Hasan T., Bhattacharjee A., Islam Md. S., Mubasshir K., Li Y., Kang Y. B., Rahman S., Shahriyar R. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4693–703. https://doi.org/10.18653/v1/2021.findings-acl.413

Hermann K. M., Kočiský T., Grefenstette E., Espeholt L., Kay W., Suleyman M., Blunsom P. Teaching Machines to Read and Comprehend. *Neural Information Processing Systems*, 2015, vol. 28, pp. 1693–1701. Available at: http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf (accessed: June 25, 2024).

Janin A., Baron D., Edwards J. A., Ellis D. P. W., Gelbart D., Morgan N., et al. The ICSI meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 Proceedings (ICASSP '03)*. 2003. Available at: https://www.researchgate.net/publication/4015071_The_ICSI_meeting_corpus (accessed: June 23, 2024).

Jin H., Yang Z., Meng D., Wang J., Tan J. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. arXiv (Cornell University). 2024. Available at: https://arxiv.org/abs/2403.02901 (accessed: June 24, 2024).

Khalman M., Zhao Y., Saleh M. ForumSum: A Multi-Speaker Conversation Summarization Dataset. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4592–4599. https://doi.org/10.18653/v1/2021.findings-emnlp.391

Koupaee M., Wang W. Y. WikiHow: A Large Scale Text Summarization Dataset. arXiv (Cornell University). 2018. Available at: https://arxiv.org/pdf/1810.09305 (accessed: June 24, 2024).

Lin C. ROUGE: A Package for Automatic Evaluation of Summaries. Text *Summarization Branches Out*. 2004. Available at: https://aclanthology.org/W04-1013.pdf (accessed: June 27, 2024).

Liu C., Wang P., Xu J., Zang L., Ye J. Automatic Dialogue Summary Generation for Customer Service. KDD '19: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 2019, pp. 1957–1965. https://doi.org/10.1145/3292500.3330683

Liu L., Lu Y., Yang M., Qu Q., Zhu J., Li H. Generative Adversarial Network for Abstractive Text Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32(1). Available at https://arxiv.org/abs/1711.09357 (accessed: June 25, 2024).

Liu Y. Fine-tune BERT for Extractive Summarization. *arXiv (Cornell University). 2019*. Available at: https://arxiv.org/pdf/1903.10318.pdf (accessed: June 23, 2024).

Luhn H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 1958, vol. 2(2), pp. 159–165. https://doi.org/10.1147/rd.22.0159

Lyu M. R., Cheng P., Li X., Balian P., Bian J., Wu Y. Automatic Summarization of Doctor-Patient Encounter Dialogues Using Large Language Model through Prompt Tuning. *arXiv (Cornell University).* 2024. Available at: https://arxiv.org/abs/2403.13089 (accessed: June 24, 2024).

Malykh V., Chernis K., Artemova E., Piontkovskaya I. SumTitles: a Summarization Dataset with Low Extractiveness. *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5718–5730. https://doi.org/10.18653/v1/2020.coling-main.503

Moratanch N., Gopalan C. A survey on Extractive Text Summarization. 2017 *International Conference on Computer, Communication and Signal Processing (ICCCSP).* 2017. Available at: https://ieeexplore.ieee.org/document/7944061 (accessed: June 25, 2024).

Napoles C., Gormley M. R., Van Durme B. Annotated Gigaword. *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX),* 2012, pp. 95–100. Available at: https://aclanthology.org/W12-3018.pdf (accessed: June 23, 2024).

Narayan S., Cohen S. B., Lapata M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 2018, pp. 1797–1807. https://doi.org/10.18653/v1/d18-1206

Nedoluzhko A., Singh M., Hledíková M., Ghosal T., Bojar O. ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3174–3182. Available at: https://aclanthology.org/2022.lrec-1.340/ (accessed: June 23, 2024).

Papineni K., Roukos S., Ward T., Zhu W. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318. Available at: https://aclanthology.org/P02-1040.pdf (accessed: June 27, 2024).

Rameshkumar R., Bailey P. Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5121–34. https://doi.org/10.18653/v1/2020.acl-main.459

Shukla A., Bhattacharya P., Poddar S., Mukherjee R., Ghosh K., Goyal P., Ghosh S. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. *arXiv (Cornell University)*, 2022. Available at: https://arxiv.org/abs/2210.07544 (accessed: June 22, 2024).

Zhang S., Çelikyılmaz A., Gao J., Bansal M. EmailSum: Abstractive Email Thread Summarization. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, vol. 1, pp. 6895–6909. https://doi.org/10.18653/v1/2021.acl-long.537

Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT. *arXiv (Cornell University)*, 2020. Available at: https://arxiv.org/pdf/1904.09675 (accessed: June 27, 2024).

Zhong M., Yin D., Yu T., Zaidi A., Mutuma M., Jha R., et al. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5905–5921. https://doi.org/10.18653/v1/2021.naacl-main.472

Zhu C., Liu Y., Mei J., Zeng M. MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5927–5934. https://doi.org/10.18653/v1/2021.naacl-main.474

Информация об авторах

Дмитриева Кристина Александровна, стажер-исследователь
Жолус Марина Романовна, стажер-исследователь, инженер-программист АО «Эврика»

Information about the Authors

Kristina A. Dmitrieva, Research Assistant
Marina R. Zholus, Research Assistant, Software Engineer

Статья поступила в редакцию 09.09.2024;
одобрена после рецензирования 21.09.2024; принята к публикации 30.09.2024

The article was submitted 09.09.2024;
approved after reviewing 21.09.2024; accepted for publication 30.09.2024