

И.Е. Лысенко

**МАЛОРЕСУРСНАЯ АБСТРАКТИВНАЯ СУММАРИЗАЦИЯ ТЕКСТОВ
С ИСПОЛЬЗОВАНИЕМ КОНТРАСТИВНОГО ОБУЧЕНИЯ БЕЗ УЧИТЕЛЯ
И ROUGE-ОРИЕНТИРОВАННОЙ ФУНКЦИИ ПОТЕРЬ**

Актуальность данной работы определяется тем, что в специализированных предметных областях и малоресурсных языках получение достаточного числа примеров «текст-реферат» для эффективного решения задачи абстрактивной суммаризации текстов оказывается дорогостоящим и практически невыполнимым, тогда как современные модели глубокого обучения требуют крупных размеченных корпусов, а большие объемы неразмеченных текстов остаются неиспользуемыми. Цель исследования заключается в разработке нового метода обучения моделей абстрактивной суммаризации для малоресурсных (10-примерных и 100-примерных) режимов, обеспечивающего рост качества за счет дообучения модели на неразмеченных данных с использованием контрастивного обучения без учителя с аугментацией входа. Из задач исследования выделяются создание новой контрастивной функции потерь и проведение сравнения полученного подхода с существующими методами малоресурсной абстрактивной суммаризации. Методы и подходы включают новую функцию потерь для дообучения трансформера, включающую в себя контрастивную генеративную компоненту дифференцируемой аппроксимации метрики ROUGE-3. Предложены два варианта метода – последовательный «DiffROUGE-seq» и метод с частичным привлечением учителя «DiffROUGE-sim». В качестве базовой модели используется BART-large, а аугментации входа генерируются моделью FLAN-T5-large в 0-примерном режиме. Произведено сравнение на популярных выборках AESLC, Gigaword, XSum и Reddit. Предложенный метод показал заметное улучшение метрик ROUGE в успешных случаях в среднем на 2,69 для 10-примерного режима и в среднем на 2,13 для 100-примерного режима. Таким образом, новый предложенный метод малоресурсной абстрактивной суммаризации, использующий как размеченные, так и неразмеченные данные, заметно более эффективен по метрике ROUGE по сравнению с существующими подходами.

Обработка естественного языка; абстрактивная суммаризация; малоресурсный режим; глубокое обучение; трансформер; контрастивное обучение; дифференцируемая аппроксимация метрики.

I.E. Lysenko

**LOW-RESOURCE ABSTRACTIVE TEXT SUMMARIZATION VIA CONTRASTIVE
UNSUPERVISED LEARNING WITH A ROUGE-ORIENTED LOSS FUNCTION**

The relevance of this work is motivated by the fact that in specialized domains and low-resource languages obtaining a sufficient number of “document-summary” pairs for effective abstractive text summarization is expensive and often practically infeasible, whereas modern deep learning models require large labeled corpora and large volumes of text remain unused. The aim of this study is to develop a new training method for abstractive summarization models in low-resource (10-shot and 100-shot) settings that improves quality by fine-tuning the model on unlabeled data using unsupervised contrastive learning with input augmentation. The research tasks include designing a new contrastive loss function and comparing the proposed approach with existing methods of low-resource abstractive summarization. The methods and approaches comprise a new loss function for fine-tuning a transformer that includes a contrastive generative component based on a differentiable approximation of the ROUGE-3 metric. Two variants of the method are proposed – the sequential “DiffROUGE-seq” and the semi-supervised “DiffROUGE-sim”. BART-large is used as the base model, while input augmentations are generated by FLAN-T5-large in a zero-shot regime. Experiments are conducted on popular datasets AESLC, Gigaword, XSum, and Reddit. The proposed method achieves substantial improvements in ROUGE scores in successful cases, with average gains of 2.69 in the 10-shot setting and 2.13 in the 100-shot setting. In summary, the newly proposed low-resource abstractive summarization method that leverages both labeled and unlabeled data is significantly more effective in terms of ROUGE than existing approaches.

Natural language processing; abstractive summarization; low-resource setting; deep learning; transformer; contrastive learning; differentiable approximation of a metric.

Введение. Стремительный рост объемов текстовой информации в науке, промышленности и медиа усложняет ее анализ и усиливает потребность в надежных методах суммаризации, которые позволяют по входному тексту получить выжимку главной информации в нем. Это позволяет решить проблему информационной перегрузки читателя

и улучшает степень восприятия содержания текстов. Различается экстрактивная суммаризация, задача которой заключается в составлении реферата путем выбора наиболее информативных предложений текста, и абстрактивная суммаризация, задача которой заключается в перефразировании и новой формулировке исходного текста в более короткой форме. На данный момент, с ростом популярности больших языковых моделей, абстрактивная суммаризация становится все более распространенной, так как она позволяет получить не только информативный, но и связный текст реферата.

Абстрактивная суммаризация, так как она основана на больших языковых моделях, которые зачастую базируются на архитектуре «трансформер» [1], требует больших объемов тренировочных пар «текст-реферат», чтобы используемая модель научилась генерировать рефераты высокого качества. Однако в специализированных областях, таких как юриспруденция, финансы или медицина, получение достаточного числа таких пар является практически невыполнимым из-за стоимости разметки, требования к специализированным знаниям и т.п., но количество неразмеченных данных в данных областях зачастую большое. Таким образом, применение абстрактивной суммаризации к текстам из таких областей существенно ограничено и является открытой проблемой.

В данной работе мы рассматриваем случай, когда имеется мало размеченных примеров «текст-реферат» (10 или 100) и много неразмеченных примеров, то есть только текстов без сопутствующих им рефератов. В отличие от существующих методов обучения больших языковых моделей, таких как BART [2], PEGASUS [3] и T5 [4], предложенный нами подход позволяет воспользоваться информацией из неразмеченного набора данных для улучшения качества абстрактивной суммаризации. Мы предлагаем новый метод обучения, который заключается в максимизации предложенной нами контрастивной модификации дифференцируемой аппроксимации популярной метрики ROUGE-3 [5] между сгенерированными моделью текстами для исходного неразмеченного текста и его аугментации. Мы также показываем, что предложенный метод зачастую оказывается заметно лучше существующих в литературе методов по метрикам ROUGE [5].

1. Описание и постановка решаемой задачи. 1.1. Формулировка проблемы. Абстрактивная суммаризация – задача обработки естественного языка, которая заключается в том, чтобы по входному тексту $X = (x_1, x_2, \dots, x_{|X|})$ получить его краткий реферат $Y = (y_1, y_2, \dots, y_{|Y|})$, такой, что он является связным текстом, охватывающим всю важную информацию из исходного текста, и $|X| \gg |Y|$.

Задача абстрактивной суммаризации обычно оценивается метрикой ROUGE-N [5]. Формально, пусть f_θ – модель абстрактивной суммаризации с параметрами θ , которая по входному тексту X порождает реферат $Y = f_\theta(X)$, и пусть $D = \{(X^{(i)}, \hat{Y}^{(i)})\}_{i=1}^{|D|}$ – тестовая выборка пар «текст – эталонный реферат». В качестве технического параметра, являющегося критерием оценки качества абстрактивной суммаризации, используется метрика ROUGE-N [5], измеряющая долю совпадающих N-грамм между сгенерированным Y и эталонным $\hat{Y}$ рефератами. Обозначив через $C_T(g)$ число вхождений N-граммы g в текст T , определим:

$$C_Y^{\hat{Y}}(g) = \min(C_Y(g), C_{\hat{Y}}(g))$$

$$p_n(Y, \hat{Y}) = \frac{\sum_{g \in Y} C_Y^{\hat{Y}}(g)}{\sum_{g \in Y} C_Y(g)}; \quad r_n(Y, \hat{Y}) = \frac{\sum_{g \in Y} C_Y^{\hat{Y}}(g)}{\sum_{g \in \hat{Y}} C_{\hat{Y}}(g)}$$

$$ROUGE_N(Y, \hat{Y}) = 100 \cdot 2 \cdot \frac{p_n(Y, \hat{Y}) \cdot r_n(Y, \hat{Y})}{p_n(Y, \hat{Y}) + r_n(Y, \hat{Y})} \in [0, 100].$$

Критерием качества модели служит усредненное по тестовой выборке значение метрики:

$$Q(\theta) = \frac{1}{|D|} \sum_{i=1}^{|D|} ROUGE_N(Y^{(i)}, \hat{Y}^{(i)}).$$

Таким образом, задача формализуется как поиск параметров θ^* , максимизирующих критерий $Q(\theta)$:

$$\theta^* = \arg \max_{\theta} Q(\theta).$$

В данном исследовании рассматривается случай абстрактивной суммаризации, когда доступно 10 или 100 пар «текст-реферат» и много неразмеченных текстов без рефератов, так как этот случай описывает существующие области текстов, например, юриспруденцию, медицину или финансы, и существующие методы абстрактивной суммаризации плохо работают с текстами из этих областей ввиду недостаточного количества размеченных данных. Таким образом, проблема исследования заключается в том, что задача абстрактивной суммаризации текстов с малым числом размеченных данных плохо решается существующими популярными методами [2–4] ввиду того, что они требуют больших объемов пар «текст-реферат» для обучения.

1.2. Актуальность темы. Актуальность темы абстрактивной суммаризации обусловлена быстрым ростом объемов текстовой информации во многих областях, которую ввиду проблемы информационного перегруза сложно просматривать, воспринимать и анализировать, поэтому возникает потребность в автоматизированных средствах сжатия и обобщения данных. Системы абстрактивной суммаризации, способные формировать связные и информативные рефераты за счет генерации новых формулировок, имеют важное значение как для развития фундаментальных исследований в области обработки естественного языка (например, как одна из задач для оценки качества новых больших языковых моделей), так и для практических приложений – от новостных агрегаторов и интеллектуальных поисковых систем до поддержки принятия решений в таких областях, как медицина, юриспруденция или финансы.

Перспективы развития темы связаны с совершенствованием методов, устойчиво работающих в режимах дефицита данных. Ожидается, что развитие подходов к абстрактивной суммаризации, основанных на эффективном использовании неразмеченных корпусов и контрастивного обучения, позволит получать более точные и содержательные резюме для широкого круга прикладных задач и областей.

1.3. Обзор литературы. **1.3.1 Методы абстрактивной суммаризации.** Одними из самых популярных моделей для решения задачи абстрактивной суммаризации на данный момент являются BART [2], PEGASUS [3] и T5 [4], основанные на архитектуре «трансформер» [1].

BART [2] – модель кодировщик-декодировщик, которая училась без учителя на задачу восстановления текста после одного из пяти возможных методов зашумления: маскировки токенов, удаления токенов, заполнения текста, перестановки предложений и циклического сдвига документа.

PEGASUS [3] – модель кодировщик-декодировщик, которая была создана для решения задачи абстрактивной суммаризации. Данная модель, аналогично BART [2], обучалась без учителя на задачу восстановления текста после зашумления, однако при обучении данной модели для зашумления исходного текста используется метод генерации пропущенных предложений, предложенный авторами, чтобы метод зашумления более тесно соответствовал целевой задаче абстрактивной суммаризации и приводил к лучшим и более быстрым результатам дообучения. Данная модель показывает хорошие результаты для малопримерных режимов и зачастую показывает наилучшие результаты среди трех рассмотренных моделей, поэтому мы используем ее в качестве базовой для сравнения с предложенным нами подходом.

T5 [4] – модель кодировщик-декодировщик, которая сначала обучалась без учителя на задачу восстановления текста по искаженным фрагментам, а затем училась на отдельные задачи (абстрактивная суммаризация, машинный перевод и т.п.), используя префикс задачи («summarize: ...», «translate English to German: ...»). Таким образом, данная модель способна решать множество различных генеративных задач одновременно, используя разные префиксы задач.

Данные модели зачастую используются в качестве базовых, вокруг которых строятся существующие подходы к малопримерной абстрактивной суммаризации.

Активное обучение – метод оптимальной разметки данных с ограниченным бюджетом, который итеративно выбирает для разметки примеры, которые имеют максимальную оценку неопределенности для данной модели. Активное обучение часто используется в случае дорогой разметки, когда размеченных данных мало, а неразмеченных – много, в частности, в абстрактивной суммаризации. IDDS [6] – метод оценки неопределенности генеративной модели для использования в активном обучении, который каждый шаг приоритизирует объекты, которые находятся в среднем далеко от размеченных данных, но в среднем близко к неразмеченным. В данной статье авторы также предлагают метод самообучения, во время которого для разметки неразмеченных данных используется сама модель, обученная на малом количестве имеющихся размеченных данных. Для экспериментов в малоресурсных режимах авторы используют модель BART [2]. Во время сравнения мы используем название «Tsvigun et al.» для обращения к лучшим подходам данной статьи.

FactPEGASUS [7] – модель абстрактивной суммаризации, основанная на модели PEGASUS [3]. Авторы, во-первых, модифицируют метод генерации пропущенных предложений PEGASUS [3], чтобы выбирать более фактические предложения, а, во-вторых, предлагают использовать во время дообучения три модуля, улучшающих фактичность итоговой модели: соединитель, добавляющий токен маскировки в резюме для соответствия задаче предобучения, корректор, убирающий галлюцинации из резюме, и контрактор, который добавляет в функцию потерь контрастивную компоненту, позволяющую модели выбирать более фактические резюме. Используя данные нововведения, FactPEGASUS [7] показывает способность к генерации более фактических резюме в малоресурсных режимах.

MTL-ABS [8] – метод малоресурсной абстрактивной суммаризации, который использует как латентные знания большой языковой модели, так и внешние корпуса текстов, для улучшения качества решения данной задачи. Авторы предлагают добавить в трансформер [1] слой-адаптеры, чтобы не обучать модель полностью и облегчить перенос знаний из внешних корпусов текстов. Для экспериментов авторы используют трансформер [1] с 12 слоями кодировщика и 6 слоями декодировщика.

UniSumm [9] – метод обучения моделей суммаризации в малоресурсном режиме. Авторы сначала тренируют суммаризационную модель на различных выборках, помечая их специальными префиксами, включая универсальный префикс для дальнейшего эффективного малопримерного обучения. Далее авторы используют универсальный префикс для дообучения модели на конкретной малопримерной области. Авторы используют BART [2] в качестве базовой суммаризационной модели.

SummaFusion-large [10] – метод абстрактивной суммаризации, который заключается в комбинации нескольких рефератов-кандидатов для получения итогового реферата. Авторы предлагают передавать в декодировщик конкатенированное векторное представление исходного текста и всех рефератов-кандидатов. Авторы утверждают, что такой подход улучшает качество в случае, когда качество отдельных кандидатов низкое, например, в малоресурсной абстрактивной суммаризации. В качестве базовой модели авторы используют PEGASUS [3].

1.3.2. Выборки данных. Мы проводим эксперименты на четырех популярных выборках данных AESLC [11], Gigaword [12], XSum [13] и Reddit [14]. Аналогично существующим методам, мы симулируем малоресурсный режим, выбирая случайные 10 или 100 размеченных тренировочных примеров для обучения с учителем и используем остальные тренировочные данные для обучения без учителя.

AESLC [11] – выборка электронных писем компании Enron, суммаризационной задачей которой является генерация темы письма по его содержанию. И тексты, и резюме данной выборки достаточно короткие – в среднем 75 и 4 слова, соответственно. Выборка содержит 14 436 тренировочных, 1 960 валидационных и 1 906 тестовых примеров. Область выборки – деловые и персональные письма.

Gigaword [12] – выборка новостного характера, задачей для которой является составление заголовка новостной статьи по ее содержанию. И тексты, и резюме данной выборки также являются короткими – в среднем 31 и 8 слов, соответственно. Выборка содержит 3 803 957 тренировочных, 189 651 валидационных и 1 951 тестовых примера.

XSum [13] – выборка статей BBC из различных областей (новости, политика, спорт, погода и т.п.), суммаризационной задачей которой является составление резюме из одного предложения по содержанию статьи. Авторы использовали подзаголовок статьи в качестве верного резюме. Тексты данной выборки являются длинными, тогда как резюме являются короткими – в среднем 431 и 23 слова, соответственно. Выборка содержит 204 045 тренировочных, 11 332 валидационных и 11 334 тестовых примера.

Reddit [14] – выборка постов социальной сети Reddit, конкретно, форума r/TIFU. Авторы используют в качестве верного резюме поле «tl;dr», которое присутствует в каждом посте и описывает его краткое содержание. Тексты данной выборки являются длинными, а резюме короткими – в среднем 432,6 и 23 слова, соответственно. Выборка содержит 33 711 тренировочных, 4 214 валидационных и 4 214 тестовых примера.

1.3.3. Литература, связанная с предложенными методами. UDA [15] – негенеративный метод контрастного обучения моделей машинного обучения, который заключается в добавлении в функцию потерь слагаемого кросс-энтропии от выходного условного вероятностного распределения модели на входном примере и от выходного условного вероятностного распределения модели на аугментации входного примера, которая сохраняет семантический смысл исходного примера, но является другим входным примером. Данный метод позволяет использовать подход с частичным привлечением учителя, приближая с точки зрения модели семантически идентичные примеры. В нашем исследовании мы также используем аналогичное контрастивное слагаемое для нашего метода «DiffROUGE-sim», однако мы используем нашу контрастивную дифференцируемую аппроксимацию метрики ROUGE-3 [5] в качестве меры близости ответов модели.

P-BLEU [16] – дифференцируемая аппроксимация метрики BLEU [17], которая часто используется для измерения качества в задаче машинного перевода и измеряет пересечение N-грамм текста, аналогично ROUGE [5]. Авторы предлагают фреймворк составления дифференцируемых аппроксимаций метрик, основанных на пересечении N-грамм, однако он требует знания верного ответа для входного текста. В нашем исследовании мы модифицируем данный фреймворк для получения контрастивной дифференцируемой аппроксимации метрики ROUGE-3 [5].

FLAN-T5 [18] – улучшенная версия модели T5 [4], которая была обучена на 1 800 различных задачах, сформулированных в виде инструкций на естественном языке. Данная модель показывает высокое качество на многих задачах в 0-примерном режиме (единственный вход модели – инструкция, без примеров решения), поэтому мы используем ее в качестве генератора аугментаций текстов в 0-примерном режиме.

ROUGE [5] – популярная метрика качества абстрактивной суммаризации, которая основана на измерении пересечения N-грамм текстов и принимает значения от 0 (отсутствие пересечения) до 100 (полное пересечение). Хотя существуют более продвинутые метрики качества абстрактивной суммаризации, например, BERTScore [19] или BARTSCORE [20], которые, как показывают исследования, лучше коррелируют с человеческим мнением, мы используем ROUGE [5] для итогового сравнения моделей, так как, во-первых, модельные метрики выше также требуют дообучения для хорошего качества на малоресурсных областях, в отличие от ROUGE [5], а, во-вторых, мы используем BART [2] в качестве базовой модели для наших подходов, и в нашем случае использовать BARTSCORE [20] некорректно, так как данная метрика будет предвзята в пользу нашей модели.

1.3.4. Используемое программное обеспечение. В качестве библиотеки для загрузки, сохранения и обучения трансформеров [1] мы используем библиотеку transformers [21], так как она является наиболее популярной для данного типа моделей. Для автоматического дифференцирования и работы с тензорами мы используем популярную библиотеку PyTorch [22]. Для загрузки, обработки и сохранения текстовых выборок данных мы используем популярную библиотеку datasets [23], так как она хорошо интегрирована с используемой нами библиотекой transformers [21].

1.4. Цель исследования. Цель исследования заключается в разработке нового метода обучения моделей абстрактивной суммаризации для малоресурсных (10-примерных и 100-примерных) режимов, обеспечивающего рост качества за счет дообучения модели на неразмеченных данных с использованием контрастивного обучения без учителя с аугментацией входа.

1.5. Задачи исследования:

1. Проведение сравнительного анализа существующих методов решения задачи малоресурсной (10-примерной и 100-примерной) абстрактивной суммаризации.
2. Создание новой контрастивной функции потерь, основанной на дифференцируемой аппроксимации метрики ROUGE-3 [5], которая позволяет использовать неразмеченные данные в малопримерном режиме обучения задачи абстрактивной суммаризации.
3. Реализация и проведение экспериментов на популярных выборках данных с предложенным методом.
4. Сравнение предложенного метода с существующими, получение оценки эффективности подхода и формулирование рекомендаций по его применению.

1.6. Объект и предмет исследования. Объектом исследования являются процессы и методы абстрактивной суммаризации текстов на основе моделей глубокого обучения. Предметом исследования является метод малоресурсного (10-примерного и 100-примерного) обучения моделей абстрактивной суммаризации, основанный на контрастивном дообучении без учителя трансформера с использованием аугментации входных данных и ROUGE-ориентированной функции потерь.

2. Методика исследования. 2.1. Стратегия и подходы к исследованию. Выделяется несколько возможных подходов решения задачи малоресурсной абстрактивной суммаризации, в частности, классическое обучение с учителем, самообучение модели, использование специализированных промптов вкупе с большими языковыми моделями и контрастивное обучение. Мы используем именно подход с контрастивным обучением, поскольку он, в отличие от других подходов, позволяет эффективно использовать неразмеченные данные, избегает псевдо-разметки и формирует инвариантность и робастность к семантически идентичным входам. Остальные подходы также имеют свои недостатки:

- ◆ Классическое обучение с учителем вызывает быстрое переобучение и имеет проблему низкого качества и высокой чувствительности к инициализации и случайности выборки.

- ◆ Самообучение модели зачастую усиливает эффект ошибок разметки данных моделью и приводит к сильному снижению качества итоговой модели.

- ◆ Большие языковые модели, во-первых, вычислительно тяжелы и дороги в использовании и дообучении, а, во-вторых, их нельзя использовать в случае, когда данные не разрешено выкладывать в открытый доступ (например, в случае соглашения о неразглашении или ограничивающего согласия об обработке данных).

Мы используем метод эксперимента для эмпирического подтверждения разумности и эффективности нашего метода по сравнению с существующими, поскольку ввиду размера и сложности используемых моделей аналитический анализ метода затруднен.

Исследование подразделяется на несколько этапов:

1. Анализ существующих методов малоресурсной абстрактивной суммаризации – позволяет выявить, какие подходы уже предложены, какие результаты они показывают и какие ограничения остаются.
2. Формирование исследовательской гипотезы – задает четкую цель, к проверке которой направлены все последующие шаги.
3. Аугментация входных данных – создает основу в виде итоговых выборок данных для дальнейшего контрастивного обучения.
4. Программная реализация предложенных методов – переводит теоретическую идею в работающий прототип.
5. Проведение экспериментов с предложенными методами – дает количественную оценку того, как новый метод работает в конкретных условиях.

6. Проведение сравнения с существующими методами малоресурсной абстрактивной суммаризации в 10-примерном и 100-примерном режимах – позволяет ответить на ключевой вопрос: лучше ли предлагаемый метод по сравнению с существующими.

7. Анализ результатов и подведение итогов – обеспечивает интерпретацию полученных численных результатов.

2.1. Описание предлагаемых методов. В качестве контрастивной метрики близости сгенерированных моделью текстов мы используем метрику ROUGE-3 [5], которая принимает значения от 0 до 100 и измеряет долю пересечений 3-грамм входных текстов. Используя существующие обозначения из литературы [16], введем формальное описание метрики ROUGE-N [5]. Для сгенерированного моделью резюме $y \in Y$ по входу $x \in X$, верного ответа $\hat{y} \in Y$ и N-граммы g :

$$C_y(g) = \sum_{t=0}^{T-n} \prod_{i=1}^n 1\{g_i = y_{t+i}\}; \quad C_{\hat{y}}(g) = \min(C_y(g), C_{\hat{y}}(g))$$

$$p_n(y, \hat{y}) = \frac{\sum_{g \in y} C_{\hat{y}}(g)}{\sum_{g \in y} C_y(g)}; \quad r_n(y, \hat{y}) = \frac{\sum_{g \in y} C_{\hat{y}}(g)}{\sum_{g \in \hat{y}} C_{\hat{y}}(g)}$$

$$ROUGE_N(y, \hat{y}) = 100 \cdot 2 \cdot \frac{p_n(y, \hat{y}) \cdot r_n(y, \hat{y})}{p_n(y, \hat{y}) + r_n(y, \hat{y})} \in [0, 100].$$

Для получения дифференцируемой аппроксимации P-ROUGE для данной метрики, основанной на полученной от языковой модели условной вероятности следующего токена, авторы [16] вводят

$$\tilde{C}_y(g) = \sum_{t=0}^{T-n} \prod_{i=1}^n 1\{g_i = y_{t+i}\} P(y_{t+i} | y_{<t+i}, x, \theta); \quad \tilde{C}_{\hat{y}}(g) = \min(\tilde{C}_y(g), C_{\hat{y}}(g)).$$

Далее,

$$\tilde{p}_n(y, \hat{y}) = \frac{\sum_{g \in y} \tilde{C}_{\hat{y}}(g)}{\sum_{g \in y} \tilde{C}_y(g)}; \quad \tilde{r}_n(y, \hat{y}) = \frac{\sum_{g \in y} \tilde{C}_{\hat{y}}(g)}{\sum_{g \in \hat{y}} \tilde{C}_{\hat{y}}(g)}$$

$$P_{ROUGE_N(y, \hat{y})} = 100 \cdot 2 \cdot \frac{\tilde{p}_n(y, \hat{y}) \cdot \tilde{r}_n(y, \hat{y})}{\tilde{p}_n(y, \hat{y}) + \tilde{r}_n(y, \hat{y})}.$$

Однако в случае обучения без учителя верный ответ $\hat{y}$ недоступен, поэтому для нашего метода мы предлагаем модифицировать P-ROUGE, чтобы, во-первых, ориентироваться только на ответы модели в отсутствие размеченных верных ответов, а, во-вторых, улучшить дифференцируемость функции для возможности использования полной информации по градиенту (в частности, функция $\min$ не является дифференцируемой в классическом смысле).

Пусть существует некоторая функция аугментации $T: X \rightarrow X$, которая для входного текста $x \in X$ преобразует его в текст $x^T \in X$, который имеет такой же семантический смысл, как и x , но $x \neq x^T$. Вместо верного ответа $\hat{y} \in Y$ используем ответ модели $y^T \in Y$ по входу x^T и используем функцию $\text{LogSumExp}(x, y) = \log(\exp(x) + \exp(y))$ для получения дифференцируемой аппроксимации функции минимума [24]. Введем предложенную нами новую контрастивную метрику близости сгенерированных моделью текстов DiffROUGE_N :

$$M = \max(\tilde{C}_y(g), \tilde{C}_{y^T}(g))$$

$$\bar{C}_y^{y^T}(g) = -\frac{1}{\beta} \text{LogSumExp}(-\beta(\tilde{C}_y(g) - M), -\beta(\tilde{C}_{y^T}(g) - M)) + M, \beta \in \mathbb{R}.$$

Таким образом,

$$\bar{p}_n(y, y^T) = \frac{\sum_{g \in y} \bar{C}_y^{y^T}(g)}{\sum_{g \in y} \bar{C}_y(g)}; \quad \bar{r}_n(y, y^T) = \frac{\sum_{g \in y} \bar{C}_y^{y^T}(g)}{\sum_{g \in y^T} \bar{C}_y(g)}$$

$$\text{DiffROUGE}_N(y, y^T) = 2 \cdot \frac{\bar{p}_n(y, y^T) \cdot \bar{r}_n(y, y^T)}{\bar{p}_n(y, y^T) + \bar{r}_n(y, y^T)}.$$

Градиент не считается по M во время обучения методом градиентного спуска, так как этот параметр предназначен только для численной стабильности метода. В наших экспериментах параметр $\beta = 20$ и используется для того, чтобы предложенная аппроксимация функции минимума была ближе к самой функции минимума. В наших экспериментах $N = 3$. Мы также не умножаем результат на 100, так как при максимизации данной функции умножение на константу не повлияет на направление градиента.

Пусть $L = \{(x_i, \hat{y}_i) \mid i = 1 \dots |L|, x_i \in X, \hat{y}_i \in Y\}$ – размеченная выборка «текст-резюме», а $U = \{x_j \mid j = 1 \dots |U|, x_j \in X\}$ – неразмеченная выборка. Пусть для обучения модели $f_\theta(x) : X \rightarrow Y$ используется функция потерь $\ell(f_\theta(x), \hat{y}), (x, \hat{y}) \in L$, для трансформеров [1] это зачастую будет кросс-энтропийная функция потерь. Тогда назовем методом обучения «DiffROUGE-seq» последовательную минимизацию ℓ в рамках обучения с учителем на ограниченной размеченной выборке и $1 - \text{DiffROUGE}_3(f_\theta(x), f_\theta(T(x))), x \in U$ в рамках обучения без учителя на большой неразмеченной выборке. Методом обучения с частичным привлечением учителя «DiffROUGE-sim» назовем минимизацию следующей контрастивной функции потерь, построенную аналогично методу UDA [15]:

$$\tilde{\ell}(x_L, \hat{y}, x_U) = \ell(f_\theta(x_L), \hat{y}) - \lambda \frac{1}{|X_U|} \sum_{x_U \in X_U} \text{DiffROUGE}_3(f_\theta(x_U), f_\theta(T(x_U))), (x_L, \hat{y}) \in L, x_U \subseteq U, \lambda \in \mathbb{R}.$$

В наших экспериментах мы используем $\lambda = 3$, $|X_U| = 5$ и $|L| = 10$ или 100. Во время дообучения модели нашим методом мы используем скорость обучения $1e-6$ и линейное расписание изменения скорости обучения с долей разогрева 0,06. В качестве базовой модели f_θ для предложенных методов используется модель BART-large [2]. В качестве T мы используем модель FLAN-T5-large [18] в 0-примерном режиме без дообучения. Обучение и аугментация данных проходят на одной видеокарте Nvidia V100 32GB VRAM.

Важно заметить, что оптимизация DiffROUGE_N в данном случае не вызывает предвзятости модели к метрике ROUGE [5] во время итоговой оценки модели, так как DiffROUGE_N используется только в качестве контрастивной метрики близости между ответами модели для разных входов для формирования у модели робастности к семантически идентичным входам, и мы не используем ее во время обучения с учителем, когда имеются размеченные верные ответы. Таким образом, формат и цель использования DiffROUGE_N фундаментально отличается от случая итоговой оценки модели, когда сравнивается ответ модели и размеченный верный ответ для одного входа.

3. Результаты исследования. Мы проводим сравнения предложенных нами методов с существующими, используя стандартную метрику качества суммаризации ROUGE [5], так как она используется повсеместно для оценки качества суммаризации, и так как она не требует дообучения моделей машинного обучения, поскольку основана на правилах.

Ниже в табл. 1 и 2 предоставлены результаты экспериментов и сравнения метрик ROUGE [5] для существующих подходов и предложенных нами «DiffROUGE-seq» и «DiffROUGE-sim» на нескольких выборках данных. **Жирным** выделен лучший результат в столбце для режима, а *курсивом* выделено второе лучшее значение метрики для методов, предложенных нами, если они лучше существующих подходов. В каждой таблице мы опускаем методы, которые не предоставляют в оригинальных статьях результаты для выборок в таблице.

Табл. 1 показывает, что на выборках AESLC [11] и Gigaword [12], в которых и тексты, и рефераты короткие, предложенные нами методы оказываются лучше существующих по всем метрикам. «DiffROUGE-seq» оказывается лучшим на AESLC [11] в 10-примерном режиме и на Gigaword [12] в 100-примерном режиме, а «DiffROUGE-sim» оказывается лучше в остальных случаях.

Таблица 1

Сравнение метрик на выборках AESLC и Gigaword

	AESLC [11]			Gigaword [12]		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
10-примерный режим						
PEGASUS-large [3]	11,97	4,91	10,84	25,32	8,88	22,55
Tsvigun et al. [6]	17,78	8,69	17,61	24,33	8,85	22,43
MTL-ABS [8]	21,27	10,79	20,85	28,98	11,86	26,74
DiffROUGE-seq	25,04	12,59	22,79	31,25	13,19	28,27
DiffROUGE-sim	23,88	11,98	22,79	31,37	13,28	28,43
100-примерный режим						
PEGASUS-large [3]	16,05	7,20	15,32	29,71	12,44	27,30
Tsvigun et al. [6]	27,31	14,52	26,69	30,16	12,32	27,90
MTL-ABS [8]	23,88	12,06	23,18	30,03	12,70	27,71
DiffROUGE-seq	30,82	15,35	29,86	33,16	14,32	30,28
DiffROUGE-sim	31,34	15,50	30,41	32,11	13,82	29,56

В табл. 2, охватывающей результаты на выборках XSum [13] и Reddit [14], в которых длинные тексты, но короткие резюме, предложенные нами методы оказываются лучше остальных во всех случаях, кроме XSum [13] в 100-примерном режиме. В случаях, когда предложенные нами методы лучше существующих, преобладает метод «DiffROUGE-seq».

Таблица 2

Сравнение метрик на выборках XSum и Reddit

	XSum [13]			Reddit [14]		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
10-примерный режим						
PEGASUS-large [3]	19,39	3,45	14,02	15,36	2,91	10,76
FactPEGASUS [7]	-	-	23,44	-	-	-
MTL-ABS [8]	-	-	-	18,03	6,41	17,10
UniSumm [9]	26,10	7,20	19,92	22,88	5,60	17,02
SummaFusion-large [10]	30,41	9,92	22,93	20,79	4,77	14,58
DiffROUGE-seq	35,20	13,16	27,21	25,69	6,74	20,03
DiffROUGE-sim	34,64	12,79	26,57	23,39	6,48	17,91
100-примерный режим						
PEGASUS-large [3]	39,07	16,44	31,27	16,64	4,09	12,92
FactPEGASUS [7]	-	-	25,44	-	-	-
MTL-ABS [8]	-	-	-	20,14	7,71	19,38
UniSumm [9]	33,33	11,36	25,85	24,54	6,17	18,30
SummaFusion-large [10]	39,86	17,01	31,68	26,09	7,51	20,22
DiffROUGE-seq	37,11	14,93	29,40	27,22	8,34	21,88
DiffROUGE-sim	37,12	14,57	28,54	26,78	7,85	20,61

Таким образом, предложенные нами контрастивные методы обучения действительно приводят к улучшению качества малоресурсной абстрактивной суммаризации почти во всех случаях, кроме XSum [13] 100-примерного режима. Мы предполагаем, что это связано со спецификой XSum [13] как задачи экстремальной абстрактивной суммаризации: рефераты из одного предложения оказываются существенно более чувствительными к аугментациям входа. В этих условиях ROUGE-ориентированная функция потерь может приводить к излишней регуляризации и смещать модель от узкоспециализированных резюме, ожидаемых в XSum [13].

Выводы. В данном исследовании мы предложили новые методы обучения генеративных языковых моделей «DiffROUGE-seq» и «DiffROUGE-sim», научная новизна которых заключается в новой функции потерь, основанной на предложенной контрастивной версии дифференцируемой аппроксимации метрики ROUGE-3 [5], которая позволяет в режиме малоресурсной абстрактивной суммаризации воспользоваться большим корпусом неразмеченных данных для улучшения качества решения задачи абстрактивной суммаризации. Практическая значимость данного исследования заключается в том, что предложенный метод позволяет значительно снизить потенциальные затраты на разметку данных для абстрактивной суммаризации в областях, в которых мало размеченных данных для этой задачи, например, в юриспруденции, финансах или медицине. Системы абстрактивной суммаризации в этих областях позволяют снизить информационный перегруз читателя и дают читателю возможность сфокусироваться только на главной информации текста. Такие системы могут быть использованы, например, для сжатия научных медицинских публикаций, финансовых отчетностей или новостей, юридических документов.

Мы провели сравнение на четырех популярных выборках данных AESLC [11], Gigaword [12], XSum [13] и Reddit [14] в 10-примерном и 100-примерном режимах и показали, что предложенные методы улучшают метрики ROUGE [5] во всех случаях, кроме XSum [13] 100-примерного режима, вероятно, из-за специфики выборки как задачи экстремальной абстрактивной суммаризации. Так как метод «DiffROUGE-seq» оказывается наилучшим в большинстве случаев, мы рекомендуем отдавать предпочтение ему, однако, если в используемых данных короткие тексты, метод «DiffROUGE-sim» может оказаться лучше. Поставленные цели и задачи исследования были достигнуты полностью. Потенциальные направления дальнейших исследований включают полную оптимизацию гиперпараметров предложенных методов (N , $|X_U|$, λ , β и т.п.), использование другой базовой модели вместо BART-large [2], а также использование дифференцируемой аппроксимации метрики ROUGE-L [5] в функции потерь.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J. Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000-6010. DOI: 10.5555/3295222.3295349.
2. Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed, A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871-7880. DOI: 10.18653/v1/2020.acl-main.703.
3. Zhang J., Zhao Y., Saleh M., Liu P. J. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization, *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 7871-7880. DOI: 10.5555/3524938.3525989.
4. Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, *The Journal of Machine Learning Research*, 2020, Vol. 21, No. 1, pp. 5485-5551. DOI: 10.5555/3455716.3455856.
5. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries, *Text Summarization Branches Out*. Association for Computational Linguistics, 2004, pp. 74-81.
6. Tsvigun A., Lysenko I., Sedashov D., Lazichny I., Damirov E., Karlov V., Belousov A., Sanochkin L., Panov M., Panchenko A., Burtsev M., Shelmanov A. Active Learning for Abstractive Text Summarization, *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, 2022, pp. 5128-5152. DOI: 10.18653/v1/2022.findings-emnlp.377.
7. Wan D., Bansal M. FactPEGASUS: Factuality-Aware Pre-training and Fine-tuning for Abstractive Summarization, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2022, pp. 1010-1028. DOI: 10.18653/v1/2022.naacl-main.74.
8. Chen, Y.-S., Shuai, H.-H. Meta-Transfer Learning for Low-Resource Abstractive Summarization, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, Vol. 35, No.14, pp. 12692-12700. DOI: 10.1609/aaai.v35i14.17503.
9. Chen Y., Liu Y., Xu R., Yang Z., Zhu C., Zeng M., Zhang Y. UniSumm and SummZoo: Unified Model and Diverse Benchmark for Few-Shot Summarization, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*. Association for Computational Linguistics, 2023, pp. 12833-12855. DOI: 10.18653/v1/2023.acl-long.718.
10. Ravaut M., Joty S., Chen N. Towards Summary Candidates Fusion, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022, pp. 8488-8504. DOI: 10.18653/v1/2022.emnlp-main.581.

11. Zhang R., Tetreault J. This Email Could Save Your Life: Introducing the Task of Email Subject Line Generation, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 446-456. DOI: 10.18653/v1/P19-1043.
12. Napoles C., Gormley M., Durme B.V. Annotated Gigaword, *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*. Association for Computational Linguistics, 2012, pp. 95-100.
13. 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*. Association for Computational Linguistics, 2018, pp. 1797-1807. DOI: 10.18653/v1/D18-1206.
14. Kim B., Kim H., Kim G. Abstractive Summarization of Reddit Posts with Multi-level Memory Networks, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 2519-2531. DOI: 10.18653/v1/N19-1260.
15. Xie Q., Dai Z., Hovy E., Luong M.-T., Le Q. V. Unsupervised data augmentation for consistency training, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. DOI: 10.5555/3495724.3496249.
16. Shao C., Chen X., Feng Y. Greedy Search with Probabilistic N-gram Matching for Neural Machine Translation, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 4778-4784. DOI: 10.18653/v1/D18-1510.
17. Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311-318.
18. Chung H.W., Hou L., Longpre S., Zoph B., Tai Y., Fedus W., Li Y., Wang X., Dehghani M., Brahma S., Webson A., Gu S.S., Dai Z., Suzgun M., Chen X., Chowdhery A., Castro-Ros A., Pellat M., Robinson K., Valter D., Narang S., Mishra G., Yu A., Zhao V., Huang Y., Dai A., Yu H., Petrov S., Chi E.H., Dean J., Devlin J., Roberts A., Zhou D., Le Q.V., Wei J. Scaling instruction-finetuned language models, *The Journal of Machine Learning Research*, 2024, Vol. 25, No. 1, pp. 3381-3433. DOI: 10.5555/3722577.3722647.
19. Tianyi Z., Varsha K., Felix W., Kilian Q. W., Yoav A. BERTScore: Evaluating Text Generation with BERT, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 311-318.
20. Yuan W., Neubig G., Liu P. BARTSCORE: evaluating generated text as text generation, *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021. DOI: 10.5555/3540261.3542349.
21. Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Le Scao T., Gugger S., Drame M., Lhoest Q., Rush A. Transformers: State-of-the-Art Natural Language Processing, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38-45. DOI: 10.18653/v1/2020.emnlp-demos.6.
22. Ansel J., Yang E., He H., Gimelshein N., Jain A., Voznesensky M., Bao B., Bell P., Berard D., Burovski E., Chauhan G., Chourdia A., Constable W., Desmaison A., DeVito Z., Ellison E., Feng W., Gong J., Gschwind M., Hirsh B., Huang S., Kalambarkar K., Kirsch, L., Lazos M., Lezcano M., Liang Y., Liang J., Lu Y., Luk CK, Maher B., Pan Y., Puhersch C., Reso M., Saroufim M., Siraichi M. Y., Suk H., Suo M., Tillet P., Wang E., Wang X., Wen W., Zhang S., Zhao X., Zhou K., Zou R., Mathews A., Chanan G., Wu P., Chintala S. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation, *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Vol. 2 (ASPLOS '24)*. ACM, 2024. DOI: 10.1145/3620665.3640366.
23. Lhoest Q., Villanova del Moral A., Jernite Y., Thakur A., von Platen P., Patil S., Chaumond J., Drame M., Plu J., Tunstall L., Davison J., Šaško M., Chhablani G., Malik B., Brandeis S., Le Scao T., Sanh V., Xu C., Patry N., McMillan-Major, J., Schmid P., Gugger S., Delangue C., Matussière T., Debut L., Bekman S., Cistac P., Goehringer T., Mustar V., Lagunas F., Rush A., Wolf T. Datasets: A Community Library for Natural Language Processing, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2021, pp. 175-184. DOI: 10.18653/v1/2021.emnlp-demo.21.
24. Boyd S., Vandenberghe L. Convex optimization. Cambridge university press, 2004.

Лысенко Иван Егорович – Финансовый университет при Правительстве Российской Федерации; e-mail: 259996@edu.fa.ru; г. Москва, Россия; кафедра искусственного интеллекта; аспирант; <https://orcid.org/0009-0003-1990-021X>.

Lysenko Ivan Egorovich – Financial University under the Government of the Russian Federation; e-mail: 259996@edu.fa.ru; Moscow, Russia; Artificial Intelligence department; postgraduate student; <https://orcid.org/0009-0003-1990-021X>.