

**Метод автоматического построения обучающих
коллекций для задачи абстрактивного аннотирования
новостных статей**

**Чернышев Даниил Иванович (Chernyshev Daniil
Ivanovich), Добров Борис Викторович
(Dobrov Boris Viktorovich)**

МГУ им. М.В.Ломоносова, Москва

*chdanorbis@yandex.ru
dobrov_bv@mail.ru*

Аннотация: Создание коллекции примеров для обучения систем абстрактивного аннотирования является затратным процессом из-за больших временных затрат и высоких требований к квалификации экспертов необходимых для написания качественной аннотаций. Для задачи абстрактивного аннотирования новостных кластеров предлагается новый метод создания коллекций для обучения нейросетевых методов аннотирования – ClusterVote, предназначенный моделировать особенности задачи путем учета информации в связанных документах. Метод может быть использован для формирования абстрактивных аннотаций различного уровня детализации, а также для получения экстрактивных аннотаций. Используя метод ClusterVote, была сформирована новая коллекция на английском и русском языках для обучения систем аннотирования новостных статей – Telegram News*CV. Экспериментальные результаты показывают, что при определенных параметрах сформированные ClusterVote коллекции имеют схожие экстрактивные характеристики с такими известными наборами данных как CNN/Daily Mail и при этом обладают более высокими показателями “фактической достоверности” – воспроизведения в

аннотациях именованных сущностей исходных текстов, а также их связей.

Ключевые слова: абстрактное аннотирование, новости, машинное обучение, нейросетевые технологии

1 Введение

Аннотирование текстов — одна из самых сложных и востребованных областей обработки естественного языка ([5], [17]). Традиционно эту задачу представляют как сжатие текста без потери смысла и решают двумя подходам: экстрактивным (англ. – extractive summarization) и абстрактным (англ. – abstractive summarization). При экстрактивном аннотировании извлекаются наиболее значимые фрагменты исходного текста. Абстрактное аннотирование допускает возможность перефразирования исходного текста, что потенциально позволяет формировать аннотации в едином стиле, в случае обзорного (многодокументного) аннотирования – формировать связные аннотации, обобщающие тексты разных документов.

В настоящее время методы абстрактного аннотирования в широких предметных областях базируются на нейросетевых подходах ([9], [14]), которым для обучения требуются большие коллекции примеров с качественными аннотациями, в идеале формируемые экспертами в предметной области. Однако, объем работы и высокие требования к квалификации экспертов, необходимые для написания качественных аннотаций в широких предметных областях, делают такой подход очень затратным. Возникает задача автоматического формирования обучающих коллекций для аннотирования с заданными и контролируемыми свойствами.

Для решения этой задачи было предложено несколько универсальных автоматических подходов построения наборов данных для аннотирования. Общая идея таких подходов – отбор наиболее значимых предложений на основе

статистических оценок и автоматическое формирование «псевдо-аннотаций», моделирующих аннотации, порождаемые экспертами. Такой подход показал себя эффективным для предобучения современных нейросетевых моделей абстрактного аннотирования, таких как Pegasus [30] и PRIMERA [28], которые показывают лучшие результаты в задаче даже при малом (меньше 1000) количестве обучающих примеров. Однако эвристические метрики, учитывающие только частотные зависимости, применяемые для оценки значимости предложений, не могут в полной мере отражать смысл исходных текстов, и поэтому псевдо-аннотации, полученные такими методами, могут значительно отличаться от сформированных экспертами. Для задачи аннотирования новостных кластеров (сообщений, посвященных описанию одного события – новостного сюжета) важно учитывать интегрированную информацию всех документов кластера.

Мы предлагаем новый метод для получения псевдо-аннотаций – ClusterVote. Идея метода ClusterVote состоит в том, что, выбирая наиболее цитируемые или упоминаемые предложения (с учетом возможного перефразирования) в кластере, мы можем получить экстрактивную аннотацию, которая бы полнее отражала усредненную точку зрения различных авторов текстов новостного сюжета.

В отличие от предыдущих подходов, наш метод способен создавать как экстрактивные, так и абстрактивные аннотации с различным уровнем детализации, учитывает структурные (структуру изложения информации) и жанрово-стилистические особенности текстов предметной области.

Используя этот подход, мы создали новый набор данных для аннотирования новостных статей Telegram News*CV на основе данных соревнования Telegram Data Clustering Contest 2020¹. Мы приводим оценку качества коллекций для

¹ <https://contest.com/data-clustering-2>

различных вариантов метода и сравниваем их эффективность с ранее предложенным подходом псевдо-аннотирования для новостей [2]. Эксперименты показывают, что при определенных значениях параметров метода ClusterVote, результирующие коллекции имеет схожие экстрактивные характеристики с такими коллекциями как CNN / Daily Mail [20], а обученные на этих коллекциях методы абстрактивного аннотирования обладают более высокой степенью “фактологической достоверности” (воспроизведения в аннотациях именованных сущностей исходных текстов, а также их связей) даже при достаточно низком уровне воспроизведения исходного текста.

2 Обзор литературы

Известно большое количество методов аннотирования [5]. Абстрактивное аннотирование получило значительный прогресс с развитием методов нейросетевого глубокого обучения [9].

2.1 Модели абстрактивного аннотирования.

На данный момент лучшие результаты в задаче абстрактивного аннотирования показывают модели Pegasus [30] и PRIMERA [28]. Качество моделей оценивается по двум основным показателям: точность воспроизведения эталонных образцов и “естественностью” текста аннотации. Первое вычисляется на основе автоматических n-граммных метрик вроде ROUGE [16], что отражает способность модели угадывать отдельные слова и фразы. Для применения модели на практике необходимо также чтобы текст аннотации был связным, соответствовал контексту исходной статьи и не противоречил известным фактам и здравому смыслу. Ранние модели абстрактивного аннотирования часто сталкивались с проблемами подмены сущностей (например, “Новая Зеландия” может заменяться на “Ирландию”) [1] и “зацикливания” процесса генерации [26], что приводило к повтору одного и того же предложения на протяжении всей

аннотации. С появлением предобученных нейросетевых языковых моделей [4], имеющим возможность учета глобального контекста, характеристика “естественности” свелась к непротиворечивости [18], для оценки которой применяют различные метрики достоверности [22].

Нейросетевые модели допускают эволюционное развитие, когда новые модели учитывают достижения предыдущих, добавляя улучшения в архитектуру нейронной сети или в методы обучения. Модель Pegasus является улучшением модели BART [15], которая в свою очередь является модификацией архитектуры BERT [4] для задачи генерации текста. BART, как и BERT, предобучен для языковых задач типа Cloze test (заполнение при обучении пропусков в тексте наиболее подходящими по контексту словами). Однако, если BERT обучен предсказывать не более одного слова в пропуске на основе построенного векторного представления входного текста, то BART рекурсивно декодирует это представление для восстановления пропущенных последовательностей произвольной длины. Благодаря такой особенности BART подходит для любых задач генерации текста, в том числе и аннотирования: на вход подается текст, где единственный пропуск обозначен сразу после текста, а модель обучается заполнять этот пропуск, сопоставляя свое предсказание с эталонной аннотацией.

Корень проблемы применения BART для задачи абстрактного аннотирования заключается в постановке задачи восстановления предложений, на которой предобучается модель. В постановке решаемой BART задачи, модель обучается восстанавливать случайные предложения, а поскольку случайное предложение может быть уникальным по содержанию и не иметь связи с остальным текстом, это приводит к тому, что модель учится приоритизировать глобальные контекстные знания и игнорировать локальный контекст. Для задачи сжатия текста, в том числе и аннотирования, сохранение локального контекста часто

важнее глобальных знаний, поэтому модель Pegasus использует специальный алгоритм отбора предложений для восстановления, который гарантирует наличие необходимой для восстановления информации в неотобранных предложениях. Благодаря такой постановке задачи предобучения модель Pegasus без учителя показывает 75% эффективности полностью обученной BART и достигает 90% менее чем за 100 примеров [30].

PRIMERA можно рассматривать как модификацию Pegasus для случая обзорного (многодокументного) аннотирования. В PRIMERA используется модификация BERT с разреженным механизмом внимания Longformer, которая позволяет обрабатывать в 4 раза больше входных слов. Также, для учета иерархии документов в задаче предобучения, предложения для восстановления в документе отбираются в соответствии с иерархией оценки значимости именованных сущностей, а в качестве входной информации используются тексты других документов. Это позволило модели PRIMERA улучшить результаты Pegasus на 15% в условиях обработки нескольких документов [28].

В настоящее время, для русского языка пока неизвестно подобных PRIMERA и Pegasus специализированных моделей. Поэтому в настоящей работе для текстов на русском языке мы рассматривали ближайшие аналоги: мультязычную версию модели mBART и модель ruT5-large, которая является адаптацией модели T5 [23] для русского языка. Модель T5 имеет ту же архитектуру, что и BART, но предобучена также на задачах восстановления порядка слов.

2.2 Наборы данных для аннотирования новостных статей.

Исторически первым набором данных для абстрактного аннотирования была коллекция New York Times Annotated Corpus [24] (более 600 тыс. пар статья-аннотация). Однако из-за ограниченного доступа, в научном сообществе для тестирования методов аннотирования большее

распространение получила коллекция CNN/Daily Mail [20] (300 тыс. пар). Но эталонные аннотации набора CNN/Daily Mail имеют высокие показатели экстрактивности (часто повторяют фрагменты исходных текстов), поэтому в 2018 году был предложен набор Xsum [21] (226 тыс. пар), который содержит аннотации длиной в одно предложение с новостного сайта BBC News. В том же году был предложен набор Newsroom [8] (1.3 млн пар), который был автоматически собран путем выделения формальных структурных фрагментов в коде интернет-страниц различных новостных изданий. Следуя этому подходу, в 2020 году был опубликован первый набор для аннотирования русскоязычных новостных статей Gazeta [11] (64 тыс. пар). В 2020 и 2021 году были опубликованы мультиязычные наборы для аннотирования новостей MLSum [25] и XLSum [12], которые также содержат статьи на русском языке. На сегодняшний день суммарное количество русскоязычных пар статья-аннотация среди всех опубликованных наборов не превышает 180 тысяч.

Главный недостаток существующих автоматически собранных наборов данных – невысокое качество эталонных аннотаций. Было показано [27], что наборы данных, собранные на основе формальной структуры веб-страниц, такие как Xsum и Newsroom содержат фрагменты, несвязанные с исходным текстом, или противоречивые аннотации. Несмотря на дополнительные усилия по фильтрации, коллекция Gazeta также содержит достаточное количество некорректных эталонных аннотаций [3].

2.3 Использование псевдо-аннотаций для машинного обучения методов аннотирования.

Наборы данных для обучения абстрактного аннотирования различаются в первую очередь жанрово-стилистическими и структурными особенностями аннотаций. Поэтому для достижения максимальной эффективности систем аннотирования желательно формировать наборы

данных под конкретные области применения. Однако привлечение экспертов для формирования новых коллекций слишком затратно. В качестве разумной альтернативы были разработаны автоматические методы построения обучающих коллекций на основе псевдо-аннотаций – аннотаций полученных автоматически на основе эвристического отбора наиболее значимых предложений текста.

Впервые эта идея была применена для обучения нейросетевой модели абстрактивного аннотирования TED [29], в которой, следуя принципу «перевернутая пирамида²», в соответствии с которым основное содержание новости излагается в начале текста, в качестве псевдо-аннотаций для новостных статей использовались первый абзацы. Но в работе [31] было показано, что такая стратегия менее эффективна для обучения, чем случайный выбор подмножества предложений, и взамен был предложен метод отбора предложений для псевдо-аннотации на основе их униграммной схожести с остальным текстом. Этот подход был позже дополнен иерархией оценки значимости именованных сущностей [28], для моделирования информационной иерархии текста.

3 Формирование коллекции псевдо-аннотаций методом ClusterVote

В качестве основы для экспериментов мы выбрали данные, предоставленные в рамках научно-технологического соревнования Telegram Data Clustering Contest 2020, которое проводила компания Telegram FZ-LLC. В рамках соревнования организатор предоставил коллекции из сотен тысяч новостей за каждый из нескольких дней на разных языках. Целью соревнования была кластеризация новостей по различным признакам: язык, категория, тема. Конечным результатом этого процесса должны были быть кластеры

² [https://en.wikipedia.org/wiki/Inverted_pyramid_\(journalism\)](https://en.wikipedia.org/wiki/Inverted_pyramid_(journalism))

новостей, которые содержат только близкие по контексту статьи.

Используя предположение о тематической близости документов в новостном кластере, мы разработали метод ClusterVote, который ранжирует предложения в соответствии с популярностью представленной информации.

3.1 Набор данных Telegram Data Clustering Contest 2020.

В наборе данных Telegram Data Clustering Contest представлены статьи на различных языках, опубликованные в течение 52 дней (ноябрь 2019, май 2020). Для исследования мы использовали только англоязычную (560 000 статей от 1346 издателей) и русскоязычную (600 000 статей от 1477 издателей) части. Telegram не предоставил разметку для данных, поэтому кластеры документов были собраны самостоятельно, используя решения победителей соревнования³. Для повышения качества кластеризации предварительно были удалены статьи длиной менее 50 слов.

Для оптимизации процесса кластеризации весь набор данных был разбит на 72-часовые сегменты с 24-часовым пересечением для учета возможной 48-часовой задержки, свойственной возможности появления обзорных новостных статей. Каждый сегмент был кластеризован алгоритмом DBSCAN [6] с косинусным расстоянием. Сначала все статьи в сегменте были кластеризованы в соответствии с именованными сущностями, а затем полученные кластеры были разбиты на подкластеры по tf-idf векторам, построенным по первым 4 предложениям, что моделировало структуру “перевернутой пирамиды”.

3.2 Метод ClusterVote.

Разработан метод ClusterVote, в рамках которого предполагается, что текстовая близость новостных документов должна отражать близость представленных в

³ <https://github.com/IlyaGusev/tgcontest>

новостях наборов фактов. Под “фактом” в настоящей работе понимается элементарная тройка <субъект-отношение-объект>, полученная на основе синтаксических правил, смоделированных признанными в научном сообществе инструментами такими как spaCy⁴ и OpenIE⁵. Следуя этой логике, предполагается, что кластеры документов образуются вокруг некоторого ядра множества фактов, которое постоянно для всех документов в кластере. Расстояние конкретного факта до этого ядра отражает насколько заинтересовано сообщество авторов и читателей в нем. Самые близкие к ядру факты (в составе ядра) были оценены большинством авторов/издателей как значимые и поэтому являются основой объективной аннотации, в то время как наиболее отдаленные были оценены авторами как несущественные для читателя и поэтому опущены во многих документах.

Учитывая, что описание фактов часто отражает основной состав предложений текста, ядро фактов кластера может быть представлено как наиболее частый набор схожих предложений. Такое множество предложений можно определить, построив отображение предложений документа в предложения других документов кластера и последующим подсчетом поддержки – количества документов, содержащих предложения, близкие по упоминаемым фактам:

$$vote(s_i) = |\{D_k | s_i \notin D_k, \exists s_j \in D_k: s_j \equiv s_i\}| \quad (3.1)$$

где D_k – документ из кластера, s_k – предложение некоторого документа.

Выбирая предложения, которые получили поддержку выше некоторого порога, можно получить аннотацию кластера с точки зрения конкретного документа. Изменение этого порога будет приводить к различной детализации аннотации. Для документа источника такая аннотация будет экстрактивной, а для других документов кластера абстрактивной. В некотором

⁴ <https://github.com/explosion/spaCy>

⁵ <https://stanfordnlp.github.io/CoreNLP/openie.html>

смысле можно считать, что документы кластера “голосуют” за предложения в других документах, что и определяет название метода ClusterVote.

Для группировки предложений использовался алгоритм DBSCAN с косинусным расстоянием и векторными представлениями, полученными с помощью SentenceTransformers из моделей paraphrase-mpnet-base-v2⁶ и sbert_large_mt_nlu_ru⁷ для английского и русского языков, соответственно. Поскольку содержание и число предложений в кластере не постоянно, порог слияния ε устанавливается динамически с помощью метода локального перебора (grid search) для гарантии, что максимальное расстояние в кластере не будет превышать некоторый порог контекстной схожести d_{par} , который был подобран как 0.2 для английского и 0.3 для русского языков.

Для экспериментов были рассмотрены два характерных варианта метода:

- CV-full – порог поддержки $t_{vote} = 1$
- CV-max – порог поддержки $t_{vote} = \max(\{vote(s_i) | s_i \in Article\})$

В каждой паре статей в качестве источника аннотации была выбрана статья с наименьшим числом общей поддержки.

Метод ClusterVote можно рассматривать как особый случай метода LexRank [10], который ранжирует предложения в соответствии с их центральностью в графе связей, но на невзвешенном графе с дополнительной фильтрацией ребер. Такая фильтрация представляет собой сведение графа близости предложений к графу кластеров, в котором единственные связные компоненты – клики. Поскольку в таком графе центральность каждой вершины не зависит от вершин за пределами клика, то LexRank точно

⁶ https://www.sbert.net/docs/pretrained_models.html

⁷ https://huggingface.co/sberbank-ai/sbert_large_mt_nlu_ru

сохранит центральность внутри клик и таким образом получит то же ранжирование, что и ClusterVote.

3.3 Метрики качества.

Для оценки качества наборов данных мы вычисляем традиционные экстрактивные метрики [8]. Покрытие экстрактивными фрагментами – это доля слов исходного текста, которые сохранились в процессе аннотирования:

$$\text{Coverage}(A, S) = \frac{1}{|S|} \sum_{f \in F(A, S)} |f| \quad (3.2)$$

где A – текст исходной статьи, S – текст аннотации и $F(A, S)$ – это набор фрагментов аннотации, которые были напрямую извлечены из текста статьи. Эта метрика показывает лексическое сходство или количество синонимичных замен.

Плотность экстрактивных фрагментов – это средняя длина экстрактивного фрагмента f к которому принадлежит каждое слово в аннотации:

$$\text{Density}(A, S) = \frac{1}{|S|} \sum_{f \in F(A, S)} \sum_{w \in f} w \cdot |f| = \frac{1}{|S|} \sum_{f \in F(A, S)} |f|^2 \quad (3.3)$$

В отличие от покрытия плотность чувствительна к количеству экстрактивных фрагментов. Например, если у нас есть аннотация, состоящая из 60 слов, и общая длина экстрактивных фрагментов составляет 42 два слова, то плотность будет 29.4 в случае только одного фрагмента и 14.7 для двух фрагментов одинаковой длины, в то время как покрытие будет 0.7 вне зависимости от разбиения. Максимальное значение плотности равно длине аннотации, поэтому для снижения эффекта разницы длин в различных наборах данных метрику удобнее нормализовать. Нормализованная плотность показывает долю многословных парафраз, возникших в процессе аннотирования.

В работе [18] было показано, что фактологические ошибки в обучающих примерах приводят к генерации недостоверных

аннотаций при применении систем абстрактного аннотирования, поэтому важно проверять их корректность. Для измерения достоверности текста не существует универсальной метрики, поскольку определение факта зависит от извлекаемого контекста. Для учета различных фактологических аспектов мы вычисляем несколько метрик.

Точность воспроизведения фраз отражает долю фраз, извлеченных из текста:

$$phrase_{acc}(A, S) = ROUGE-2_{prec}(S, A) \quad (3.4)$$

Высокий уровень экстрактивности фраз гарантирует, что предложения аннотации соответствуют содержанию исходного текста, в то время как низкий показывает чрезмерное перефразирование или наличие внешней информации.

Точность воспроизведения именованных сущностей – это доля именованных сущностей правильно воспроизведенных в аннотации:

$$NEO(A, S) = \frac{|NE(A) \cap NE(S)|}{|NE(S)|} \quad (3.5)$$

где $NE(T)$ – именованные сущности текста T . Именованные сущности определяют контекстное ядро текста и поэтому в большинстве случаев не допускают перефразирования. Точность представленных фактов – это доля троек <субъект-отношение-объект>, извлеченных из аннотации, которые поддерживаются исходным текстом:

$$fact_{acc}(A, S) = \frac{|Facts(A) \cap Facts(S)|}{|Facts(S)|} \quad (3.6)$$

$fact_{acc}$ – это основной способ вычисления фактологической точности в аннотировании [7]. Для извлечения именованных сущностей и троек фактов для метрик мы использовали spaCy и OpenIE.

Основная проблема упомянутых метрик – они чувствительные к синонимичным заменам. Это делает их недостоверными в условиях частого перефразирования. Для учета таких случаев мы дополнительно измеряем точность по BERTScore [31], которая показала высокую корреляцию с мнением экспертов в сравнительном тестировании фактологических метрик FRANK⁸ [22].

3.4 Статистика для англоязычной части коллекции.

Применив метод ClusterVote к построенным ранее кластерам, мы получили 110 713 пар статья-аннотация для англоязычной части. Эту часть набора данных мы обозначили как Telegram News*CV(EN)⁹.

Для оценки естественности псевдо-аннотаций мы проводим сравнение с существующими наборами с точки зрения метрик качества: CNN/Daily Mail и Newsroom. CNN/Daily Mail является стандартным набором для обучения и тестирования систем аннотирования, а Newsroom является ближайшим аналогом нашему методу, поскольку использует автоматический подход извлечения аннотации, выделяя маркеры эталонных аннотаций из HTML кода интернет-страницы. В дополнение мы также сравниваем аннотации полученные ClusterVote с ранее предложенным способом [2] построения абстрактных псевдо-аннотаций на основе первого абзаца из статьи-аналога (обозначен как Pseudo Lead).

Примеры сравниваемых стратегий построения псевдо-аннотаций приведены в Таблице 1. Выделенные в аннотации CV-full фрагменты пересекаются с CV-max, а подчеркнутые – с Pseudo Lead.

В Таблице 2 приводится статистика по экстрактивным характеристикам наборов данных. Как можно заметить, тексты статей в нашем наборе данных значительно короче по

⁸ <https://frank-benchmark.herokuapp.com/>

⁹ Код метода ClusterVote для формирования коллекции Telegram News*CV будет выложен в репозитории <https://github.com/dciresearch/ClusterVote>

сравнению с коллекциями CNN/DM и Newsroom. Тем не менее, несмотря на высокие показатели экстрактивного покрытия (coverage), нормализованная экстрактивная плотность у всех вариантов псевдо-аннотаций ниже, чем у Newsroom, причем самое низкое значение наблюдается у варианта CV-full. Поскольку CV-max является подмножеством CV-full, то такой результат говорит о том, что предложения с меньшей поддержкой имеют большую долю перефразированные фрагментов.

Табл. 1. Примеры различных стратегий построения псевдо-аннотаций для английского языка

Источник аннотации:	https://www.vanguardngr.com/2020/04/ronaldinho-says-arrest-and-confinement-has-been-hard/
Pseudo Lead:	Kindly Share This Story: <u>Former Brazilian football star Ronaldinho said on Monday that his arrest and subsequent house arrest in Paraguay for using a false passport was “a very hard blow”.</u> “I would never have imagined myself in such a situation,” said the former Barcelona and Paris Saint-Germain striker in an interview with the Paraguayan newspaper ABC.
CV-full:	<u>Former Brazilian football star Ronaldinho said on Monday that his arrest and subsequent house arrest in Paraguay for using a false passport was “a very hard blow”.</u> Ronaldinho and his brother Roberto de Assis Moreira are accused of entering Paraguay in possession of false passports. “I was completely caught off guard when I found out that these passports were not valid,” Ronaldinho said in his first public statement since his arrest two days after arriving in Asuncion at the beginning of March. Since April 7 they have been under house arrest at the Palmaroga Hotel in the historic centre of the Paraguayan capital. Ronaldinho said he hopes to be released “as soon as possible” after cooperating with the Paraguayan police investigation. Ronaldinho and his brother face up to five years in prison if found guilty.
CV-max:	Ronaldinho and his brother Roberto de Assis Moreira are accused of entering Paraguay in possession of false passports. Since April 7 they have been under house arrest at the Palmira Hotel in the historic centre of the Paraguayan capital. Ronaldinho and his brother face up to five years in prison if found guilty.

В Таблице 3 указана статистика по фактологическим метрикам. Для всех вариантов псевдо-аннотаций экстрактивные фактологические метрики выше, чем для аннотаций из популярных наборов данных, однако вариант

Pseudo Lead имеет самые низкие показатели достоверности и чаще всех допускает ошибки при воспроизведении именованных сущностей.

Табл. 2. Сравнение наборов данных для английского языка

Набор данных	Статья # слов	Аннотаци я # слов	Coverage	Density	
				raw	norm.
CNN/DM	781	56	89%	3,87	0,07
Newsroom	659	27	83%	9,51	0,36
Telegram News*CV (EN)					
Pseudo Lead	438	88	87%	26,10	0,29
CV-full		237	91%	43,43	0,18
CV-max		95	92%	28,81	0,30

Низкие показатели *fact_{acc}* в Таблице 3 для CNN/Daily Mail связаны с самыми низкими экстрактивными характеристиками, поскольку метрика считает за ошибку несоответствие любой из компонент тройки факта <субъект-отношение-объект>. Поэтому и необходима метрика BERTScore. С точки зрения этой метрики CNN/Daily Mail имеет тот же уровень достоверности, что и Newsroom, однако, среди псевдо-аннотаций самым достоверным вариантом становится CV-full. Учитывая более низкие показатели экстрактивности по сравнению с CV-max, этот факт вновь указывает на наличие значительного перефразирования в предложениях с более низкой поддержкой.

Табл. 3. Фактологические метрики для англоязычных наборов данных

Набор данных	<i>phrase_{acc}</i>	NEO	<i>fact_{acc}</i>	BERTScore
CNN/Daily Mail	49,78%	78,12%	9,39%	36,28%
Newsroom	53,89%	76,42%	38,39%	38,42%
Telegram News*CV (EN)				
Pseudo Lead	72,10%	75,53%	44,57%	63,41%
CV-Full	79,28%	80,18%	54,57%	69,77%
CV-Max	83,09%	84,38%	61,90%	63,15%

3.5 Статистика для русскоязычной части коллекции.

Для русского языка – коллекция Telegram News*CV(RU) – получилось 173 482 пар статья-аннотация. Для сравнения мы взяли набор Gazeta, поскольку он является единственным качественным набором для русского языка. Также, как и для английского мы приводим сравнение со стратегией псевдо-аннотирования Pseudo Lead. Примеры приведены в Таблице 4.

Поскольку рассматриваемые метрики были разработаны без учета флективности русского языка, при вычислении метрик использовались лемматизированные версии текстов.

Табл. 4. Примеры различных стратегий построения псевдо-аннотаций для русского языка

Источник	https://biwork.ru/news/koronavirus-v-altajskom-krae-dannye-na-2000-6-maa
аннотации:	
Pseudo Lead:	<u>Информация по Алтайскому краю готовится на основе сводок Оперштаба края.</u> На два района расширилась география коронавируса в Алтайском крае. За минувшие сутки в Алтайском крае зарегистрировано 33 новых случая заболевания коронавирусной инфекцией.
CV-full:	<u>Информация по Алтайскому краю готовится на основе сводок Оперштаба края.</u> За минувшие сутки в Алтайском крае зарегистрировано 33 новых случая заболевания коронавирусной инфекцией. Среди новых случаев заболевания: жители Барнаула – 14 человек, Змеиногорского района – 9, Рубцовска – 2, Бийска – 1, Немецкого национального района – 1, Тальменского района – 1. Впервые зарегистрировано заболевание у трех жителей Ребрихинского и двух жителей Благовещенского районов. Из общего числа заболевших 56 случаев – завозных, в том числе 2 – из-за рубежа; 430 случаев – по контакту с заболевшими. 84 – выявлены у больных внебольничной пневмонией, 94 – у заболевших ОРВИ, в 410 случаях – бессимптомное течение. <u>Общее количество заболевших в регионе составляет 588 человек, 110 из них выздоровели и выписаны из клиник края, в том числе 2 пациента - за прошедшие сутки.</u>
CV-max:	За минувшие сутки в Алтайском крае зарегистрировано 33 новых случая заболевания коронавирусной инфекцией. <u>Общее количество заболевших в регионе составляет 588 человек, 110 из них выздоровели и выписаны из клиник края, в том числе 2 пациента - за прошедшие сутки.</u>

Статистика по наборам представлена в Таблице 5. Можно заметить, что по своим характеристикам набор Gazeta схож

больше с CNN / Daily Mail чем с Newsroom, однако степень экстрактивности ещё ниже (нормализованная плотность экстрактивных фрагментов всего 0,05). Если сравнивать псевдо-аннотации, то они также более абстрактивные, причем покрытие экстрактивными фрагментами даже ниже, чем для Gazeta. Псевдо-аннотации CV-max самые экстрактивные в данном случае, но разница с остальными вариантами пренебрежительно мала.

Возникает вопрос, на сколько будут эффективные фактологические метрики, если все варианты имеют еще меньшую степень экстрактивности чем аннотации из набора CNN / Daily Mail.

Табл. 5. Сравнение наборов данных для русского языка

Набор данных	Статья # слов	Аннотаци я # слов	Coverage	Density	
				raw	norm.
Gazeta	750	53	83%	2,46	0,05
Telegram News*CV (RU)					
Pseudo Lead	215	56	76%	6,77	0,12
CV-full		107	78%	11,19	0,10
CV-max		41	80%	7,41	0,18

В Таблице 6 представлены значения. Метрика $fact_{acc}$ не может быть посчитана для русского языка в силу отсутствия стандартного (общедоступного) инструмента для извлечения троек фактов и поэтому отсутствует в таблице. Значение метрики $phrase_{acc}$ значительно ниже в силу свободного порядка слов в русском языке, который допускает перестановки слов, не сохраняющие биграммы. С точки зрения NEO характеристики русскоязычных наборов эквивалентны англоязычным аналогам, но в данном случае CV-full оказывается наименее достоверным, находясь на том же уровне, что и Pseudo Lead. При этом BERTScore для всех выше, для Gazeta он составляет 75%, а для псевдо-аннотаций в районе 80%, с максимальным значением 81% также для CV-full.

4 Оценка качества псевдо-аннотаций

Пригодность использования набора данных для задачи аннотирования мы будем оценивать путем проведения экспериментов с существующими системами в традиционной постановке, когда часть коллекции используется для обучения, а оставшаяся часть для тестирования. При этом, если обученный на коллекции метод воспроизводит именованные сущности и фактологические взаимосвязи между ними, то такая коллекция признается приемлемой.

Табл. 6. Фактологические метрики для русскоязычных наборов данных

Набор данных	<i>phrase_{acc}</i>	NEO	BERTScore
Gazeta	15,69%	78,40%	75,41%
Telegram News*CV (RU)			
Pseudo Lead	23,84%	76,15%	79,02%
CV-Full	31,54%	75,88%	81,03%
CV-Max	21,27%	83,21%	80,14%

Более подробно – низкие показатели эффективности методов свидетельствует о том, что набор данных содержит некорректные примеры [13], которые смещают статистические характеристики задачи, делая невозможным нахождение оптимального решения, в то время как высокие говорят о тривиальности стратегии решения задачи [21], из чего следует, что обученные на таком наборе данных системы не будут способны адаптироваться к неидеальным условиям применения на практике. При этом по мере улучшения основных показателей качества аннотирования во время обучения показатели фактологической достоверности должны не уменьшаться. В противном случае эталонные аннотации не удовлетворяют главному ограничению задачи – сохранению целостности ключевой информации [18].

4.1 Базовые модели экстрактивного аннотирования.

Во многих работах по аннотированию приводится сравнение с тремя базовыми методами: Lead-k, Oracle и TextRank.

Lead-k представляет собой простое извлечение первых k предложений исходного текста, что моделирует схему “перевернутая пирамида”.

Oracle – является способом оценки верхней границы эффективности. Идея заключается в последовательном отборе предложений исходного текста жадным алгоритмом, который отбирает предложения до тех пор, пока растет n -граммная схожесть с эталонной аннотацией.

TextRank [19] один из первых способов построения экстрактивной аннотаций, использующий тот же алгоритм что и LexRank, но предназначенный для аннотирования одного документа.

4.2 Параметры экспериментов.

В модели Oracle максимизируется сумма ROUGE-1 и ROUGE-2. В модели Lead-k параметр $k = 3$ для Pseudo Lead и CV-max, и $k = 6$ для CV-full. В методе TextRank размер аннотации равен среднему размеру аннотаций в наборе данных. Для чистоты эксперимента все нейросетевые модели абстрактного аннотирования получают на вход только первые 1024 токена исходного текста статьи, что является пределом для всех моделей, кроме PRIMERA. Для генерации аннотаций используется алгоритм beam search с размером луча равным 5 и блокировкой триграмм. В качестве метрик качества измеряются ROUGE F1 и фактологические метрики $phrase_{acc}$, NEO, $fact_{acc}$, и BERTScore. Для русского языка метрики рассчитывались на лемматизированных версиях текстов.

4.3 Результаты для Telegram News*CV(EN).

В Таблице 7 приведены результаты для стратегии построения псевдо-аннотаций Pseudo-Lead. Как и ожидалось, для этого случая простое извлечение первых 3 предложений оказывается достаточным оптимальным решением. Однако разница с Oracle составляет всего 6%, что указывает на тривиальность такого варианта псевдо-аннотации. Поскольку

настоящая аннотация обычно отличается от первых 3 предложений, то предобученные модели уступают тривиальной схеме даже после полного обучения.

Табл. 7. Результаты для стратегии Pseudo Lead на Telegram News*CV(EN)

Модель	ROUGE-1	ROUGE-2	ROUGE-L	# слов
Экстрактивные				
Lead-3	74,41%	64,31%	68,57%	84
Oracle	80,16%	70,67%	74,40%	85
TextRank	47,46%	30,89%	37,81%	92
Абстрактные (Без обучения)				
PRIMERA	36,49%	19,33%	25,63%	146
Pegasus	42,68%	28,23%	34,62%	79
Абстрактные (Полное обучение)				
PRIMERA	72,13%	62,66%	66,43%	102
Pegasus	76,17%	66,53%	70,68%	90

В Таблице 8 представлены результаты для стратегии CV-full. Несмотря на более длинные аннотации, верхняя граница эффективности систем аннотирования выше, чем в случае Pseudo Lead. Тривиальная стратегия Lead-6 уже не такая эффективная, а обученные абстрактные модели значительно превосходят её, хотя и существенно отстают от метода Oracle.

В Таблице 9 отображены результаты для стратегии CV-max. Поскольку эта стратегия использует подмножество предложений CV-full, то результаты сопоставимы. Эффективность методов без учителя гораздо ниже, Lead-3 по ROUGE-2 (покрытие биграмм/фраз) всего 38%, а абстрактные модели без обучения показывают не более 24%. При этом показатели Oracle для обоих вариантов ClusterVote по этой же метрике на 10% выше, чем для Pseudo Lead, что говорит о том, что аннотации представляют собой комбинацию фрагментов исходного текста.

Табл. 8. Результаты для стратегии CV-full
на Telegram News*CV(EN)

Модель	ROUGE-1	ROUGE-2	ROUGE-L	# слов
Экстрактивные				
Lead-6	56,96%	47,49%	50,24%	176
Oracle	86,15%	78,70%	80,87%	227
TextRank	58,28%	44,60%	47,77%	218
Абстрактивные (Без обучения)				
PRIMERA	45,60%	30,25%	32,81%	146
Pegasus	38,00%	28,01%	31,95%	79
Абстрактивные (Полное обучение)				
PRIMERA	66,91%	57,24%	60,02%	162
Pegasus	65,47%	56,29%	58,80%	160

Табл. 9. Результаты для стратегии CV-max
на Telegram News*CV(EN)

Модель	ROUGE-1	ROUGE-2	ROUGE-L	# слов
Экстрактивные				
Lead-3	52,07%	38,60%	43,56%	84
Oracle	87,00%	81,14%	83,72%	85
TextRank	43,04%	26,73%	33,70%	92
Абстрактивные (Без обучения)				
PRIMERA	37,03%	21,60%	27,32%	146
Pegasus	39,12%	23,71%	30,87%	79
Абстрактивные (Полное обучение)				
PRIMERA	60,92%	51,36%	54,87%	114
Pegasus	64,91%	55,33%	59,09%	88

Распределения схем извлечения предложений методом Oracle изображено на Рис. 1. Для Pseudo Lead распределение смещено в сторону первых 3 предложений, что объясняет низкую разницу между Oracle и Lead-3 для этого типа аннотаций. Для CV-full, наоборот, распределение практически равномерное для первой половины документа с постепенным затуханием вероятности извлечения для последних предложений. CV-max будучи подмножеством CV-full лишь незначительно сдвигает распределение, но в сторону распределения набора CNN/Daily Mail. Поскольку CNN/Daily Mail содержит аннотации, написанные человеком, это говорит

о том, что CV-max отражает общий замысел экспертов, сохраняя логику повествования перевернутой пирамиды.

Для оценки корректности псевдо-аннотаций мы сравниваем результаты полученные нейросетевыми моделям в Таблице 10.

Если примеры обучающей коллекции содержали существенные фактологические несоответствия, то для их воспроизведения модели должны были обучаться “угадывать” правильные варианты и тем самым нарушать достоверность аннотации.

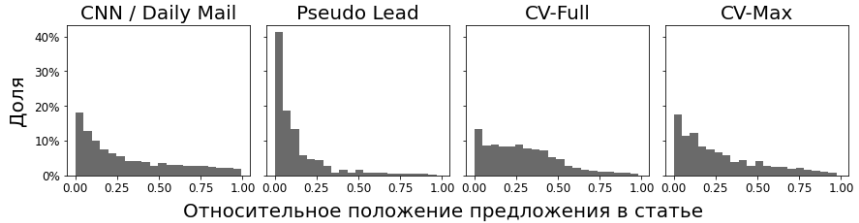


Рис. 1. Сравнение распределений относительных позиций предложений, извлеченных методом Oracle для англоязычной части

Табл. 10. Фактологические метрики для англоязычных систем аннотирования

Модель	$phrase_{acc}$	NEO	$fact_{acc}$	BERTScore
Без обучения				
PRIMERA	86,74%	90,69%	76,47%	54,65%
Pegasus	98,13%	98,74%	93,42%	70,16%
Обучение – Pseudo Lead				
PRIMERA	87,97%	85,74%	64,83%	73,10%
Pegasus	93,74%	90,12%	74,82%	74,62%
Обучение – CV-full				
PRIMERA	91,61%	88,64%	69,76%	79,78%
Pegasus	95,99%	93,26%	80,35%	84,58%
Обучение – CV-max				
PRIMERA	96,02%	95,04%	82,51%	75,95%
Pegasus	96,44%	95,49%	83,10%	79,71%

Как можно видеть, фактологические метрики сгенерированных аннотаций для всех случаев выше, чем для

псевдо-аннотаций, на которых модели обучались. Однако в случае Pseudo Lead наблюдается падение всех метрик после обучения для обеих моделей. Учитывая, что именованные сущности не должны перефразироваться в аннотации, получается, что Pseudo Lead создает недостоверные эталоны. При CV-full NEO тоже падает, но не так существенно и при этом рост BERTScore в этом случае гораздо сильнее, что говорит о большей контекстной уместности этих ошибок. Учитывая, что при CV-max наблюдается рост по всем метрикам, то скорее всего среди предложений с низкой поддержкой, которые были включены в CV-full, есть недостоверные.

4.4 Результаты для Telegram News*CV(RU).

В Таблице 11 приведены результаты стратегии Pseudo Lead для русского языка. Как и для случая английского языка, Lead-3 является оптимальным решением и отличается от Oracle всего на 6% по ROUGE-1. Но при этом верхняя граница по Oracle составляет 41%, что говорит либо о высокой абстрактности аннотаций, либо о наличии в них несоответствий с текстом статьи.

Табл. 11. Результаты для стратегии Pseudo Lead на Telegram News*CV(RU)

Модель	ROUGE-1	ROUGE-2	ROUGE-L	# слов
Экстрактивные				
Lead-3	34,42%	15,58%	33,32%	59
Oracle	41,23%	20,32%	39,74%	53
TextRank	22,57%	9,25%	21,71%	71
Абстрактные (Полное обучение)				
mBART	32,99%	15,24%	31,98%	51
ruT5	34,28%	16,18%	33,35%	56

Как можно видеть в Таблице 12, для CV-full более высокие экстрактивные показатели. Верхний порог по ROUGE-1 составляет 50%, а разница с Lead-6 уже 14%. Модели

абстрактного аннотирования уже значительно превосходят по качеству тривиальное решение.

В Таблице 13 наблюдается подобное поведение моделей для CV-max, но экстрактивность псевдо-аннотаций ниже, чем у Pseudo Lead. Стоит отметить, что в этом случае наблюдается одинаковая эффективность нейросетевых моделей абстрактного аннотирования, что говорит об устойчивости по данным оптимальной стратегии решения.

На Рис. 2 изображено распределение стратегии Oracle. Несмотря на схожесть Gazeta с CNN/Daily Mail, оптимальная стратегия отказывается смещена в сторону первого предложения.

Табл. 12. Результаты для стратегии CV-full
на Telegram News*CV(RU)

Модель	ROUGE-1	ROUGE-2	ROUGE-L	# слов
Экстрактивные				
Lead-6	36,52%	18,18%	34,68%	116
Oracle	50,16%	29,86%	48,20%	98
TextRank	34,95%	17,21%	33,23%	124
Абстрактные (Полное обучение)				
mBART	37,98%	15,98%	36,29%	102
ruT5	39,76%	21,50%	38,29%	109

Подобное распределение наблюдается для Pseudo Lead, однако вероятность извлечения предложений резко падает уже после первой четверти. Для CV-full и CV-max наблюдается мультимодальность, причем более строгие критерии отбора CV-max практически не изменяют распределение, что говорит о равнозначимости позиции в тексте при определении степени важности для русского языка.

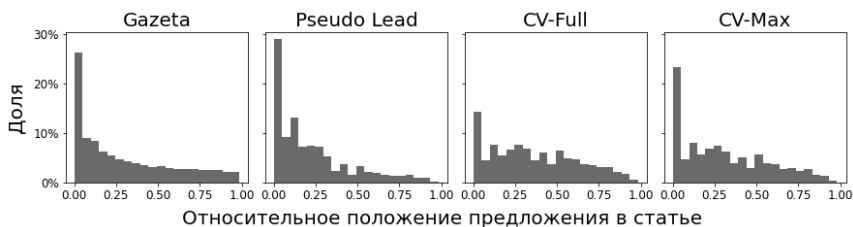


Рис. 2. Сравнение распределений относительных позиций предложений, извлеченных методом Oracle для Telegram News*CV(RU)

В Таблице 14 приведены фактологические метрики. Из-за низкой степени экстрактивности псевдо-аннотаций $phrase_{acc}$ сохраняется низким после обучения.

Табл. 13. Результаты для стратегии CV-max на Telegram News*CV(RU)

Модель	ROUGE-1	ROUGE-2	ROUGE-L	# слов
Экстрактивные				
Lead-3	21,13%	8,19%	20,43%	59
Oracle	36,77%	18,95%	35,90%	43
TextRank	17,57%	7,21%	17,15%	47
Абстрактные (Полное обучение)				
mBART	26,74%	13,29%	26,36%	42
ruT5	26,96%	12,48%	26,38%	44

Табл. 14. Фактологические метрики для русскоязычных систем аннотирования

Модель	$phrase_{acc}$	NEO	BERTScore
Обучение – Pseudo Lead			
mBART	37,07%	65,57%	81,97%
ruT5	31,27%	61,75%	80,16%
Обучение – CV-full			
mBART	36,86%	64,66%	79,60%
ruT5	44,22%	65,49%	83,26%
Обучение – CV-max			
mBART	28,30%	69,77%	81,47%
ruT5	28,19%	72,12%	81,50%

При этом показатели BERTScore высокие во всех случаях. Минимальный NEO наблюдается для моделей обученных на Pseudo Lead, а максимальный при обучении на CV-max несмотря на то, что $phrase_{acc}$ на 10% ниже для последнего, хотя разница самих аннотаций по этой метрике в наборе данных была менее 2%. Учитывая, что достоверность аннотаций модели ruT5 обученной на Pseudo Lead существенно ниже, чем для mBART, а значения ROUGE выше, то можно утверждать, что псевдо-аннотации этого типа содержат неуместную информацию и непригодны для обучения моделей абстрактного аннотирования. В случае CV-full и CV-max не наблюдается такое противоречие в метриках, но рост NEO при падении $phrase_{acc}$ в случае CV-max говорит о том, что парафразы и прямые цитаты одинаково оцениваются алгоритмом ClusterVote для русского языка, а также о большей поддержке фактологически правильных предложений.

5 Заключение

Предложен новый метод ClusterVote для автоматического построения обучающей коллекции для задачи аннотирования новостных документов, который в отличие от предыдущих подходов создает псевдо-аннотации с учетом связей с другими документами кластера. Используя этот метод, мы создали новую коллекцию Telegram News*CV, которая включает аннотации для англоязычных и русскоязычных новостных статей. Мы протестировали два варианта ClusterVote: CV-full – предложения, которые получили поддержку хотя бы от одного документа кластера, CV-max – предложения с максимальной поддержкой в кластере. Мы сравнили эти подходы с нашим предыдущим методом Pseudo Lead, который использует особенности новостного стиля “перевернутая пирамида” и статьи-аналоги.

Согласно статистике на основе рассмотренных метрик, вариант CV-max формирует самые экстрактивные и

фактически достоверные аннотации. Для английского языка этот тип псевдо-аннотации оказывается наиболее приближенным к коллекциям, полученным с привлечением экспертов, показывая то же распределение оптимальной экстрактивной стратегии, что и эталонный набор данных CNN/Daily Mail. Для русского языка это распределение сильнее смещено в сторону первых предложений даже для оригинальных аннотаций из набора Gazeta, но для псевдо-аннотаций метода ClusterVote этот эффект значительно меньше.

Результаты экспериментов показывают, что псевдо-аннотации ClusterVote, в отличие от псевдо-аннотаций Pseudo Lead, учат нейросетевые модели абстрактного аннотирования сохранять фактологическую достоверность. При этом самые устойчивые и достоверные результаты дают модели, обученные на варианте CV-max. CV-full незначительно уступает по среднему качеству, но превосходит по показателям BERTScore, что говорит о пригодности для обучения моделей любых промежуточных вариантов детализации псевдо-аннотации ClusterVote.

6 Финансирование

Работа Д.И. Чернышева (проведение экспериментов, реализация методов, обзор литературы) выполнена при финансовой поддержке Некоммерческого Фонда развития науки и образования “Интеллект”. Работа Б.В. Доброва (общая идея, анализ и интерпретация результатов) выполнена при финансовой поддержке РНФ (проект № 21-71-30003).

7 Конфликт интересов

Процесс написания и содержание статьи не дает оснований для постановки вопроса о конфликте интересов.

8 Соответствие этическим стандартам

Данная статья является полностью оригинальным произведением ее авторов, прежде не опубликована и до получения решения редколлегии PRIA о непринятии ее к публикации не будет направляться в другие издания.

9 Список литературы

1. Cao Z., Wei F., Li W., and Li S. Faithful to the original: fact-aware neural abstractive summarization. // Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). - 2018.
2. Chernyshev D. and Dobrov B. Abstractive Summarization of Russian News Learning on Quality Media // Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science. - 2021.
3. Chernyshev D. and Dobrov B. Improving Neural Abstractive Summarization with Reliable Sentence Sampling. // Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2021. Communications in Computer and Information Science, vol 1620. - 2022.
4. Devlin J., Chang M.-W., Lee K. and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. - 2019.
5. El-Kassas W.S., Salama C.R., Rafea A.A. and Mohamed H.K. Automatic text summarization: A comprehensive survey // Expert Systems with Applications, 165, 113679 - 2021.
6. Ester M., Kriegel H., Sander J. and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International

Conference on Knowledge Discovery and Data Mining (KDD'96). - 1996.

7. Goodrich B., Rao V., Liu P. J. and Saleh M. Assessing The Factual Accuracy of Generated Text // KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - 2019.
8. Grusky M., Naaman M. and Artzi Y. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. - 2018.
9. Gupta S. and Gupta S.K. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121, 49-65. - 2019.
10. Günes E. and Radev D. R. LexRank: graph-based lexical centrality as salience in text summarization // Journal of Artificial Intelligence Research. 2004.
11. Gusev I., Dataset for Automatic Summarization of Russian News // arXiv:2006.11063. - 2020.
12. Hasan T., Bhattacharjee A., Islam M. S. et al. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. - 2021.
13. Kang D. and Hashimoto T. Improved Natural Language Generation via Loss Truncation // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. - 2020.
14. Kryściński W., Rajani N. F., Agarwal D., et al. BookSum: A Collection of Datasets for Long-form Narrative Summarization // arXiv:2105.08209. - 2021.
15. Lewis M., Liu Y., Goyal N. et al. 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.

16. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out // Association for Computational Linguistics. - 2004.
17. Ma C., Zhang W. E. Guo M. et al. Multi-document summarization via deep learning techniques: A survey. ACM Computing Surveys (CSUR) - 2020.
18. Maynez J., Narayan S., Bohnet B. and McDonald R. On Faithfulness and Factuality in Abstractive Summarization // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. - 2020.
19. Mihalcea R. and Tarau P. TextRank: Bringing Order into Texts // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. - 2004.
20. Nallapati R., Zhou B., Gulcehre C. and Xiang B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond // Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. - 2016.
21. Narayan S., Cohen S. B. and 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.
22. Pagnoni A., Balachandran V. and Tsvetkov Y. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. - 2021.
23. Raffel C., Shazeer N., Roberts A. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research. - 2019.
24. Sandhaus E., The New York Times Annotated Corpus // Linguistic Data Consortium. - 2008.
25. Scialo T., Dray P.-A., Lamprier S. et al. MLSUM: The Multilingual Summarization Corpus // Proceedings of the 2020

- Conference on Empirical Methods in Natural Language Processing (EMNLP). - 2020.
26. See A., Liu P. J. and Manning C. D. Get To The Point: Summarization with Pointer-Generator Networks // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). - 2017.
 27. Tejaswin P., Naik D. and Liu P. How well do you know your summarization datasets? // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. - 2021.
 28. Xiao W., Beltagy I., Carenini G. and Cohan A. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). - 2022.
 29. Yang Z., Zhu C., Gmyr R. et al. TED: A Pretrained Unsupervised Summarization Model with Theme Modeling and Denoising // Findings of the Association for Computational Linguistics: EMNLP 2020. - 2020.
 30. Zhang J., Zhao Y., Saleh M. and Liu P. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization // Proceedings of the 37th International Conference on Machine Learning. - 2020.
 31. Zhang T., Kishore V., Wu F. et al. BERTScore: Evaluating Text Generation with BERT // International Conference on Learning Representations. - 2020.

10. Краткие биографии и фотографии всех авторов



Чернышев Даниил Иванович, аспирант кафедры алгоритмических языков факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, стажер-исследователь Научно-исследовательского вычислительного центра МГУ имени М.В. Ломоносова.



Добров Борис Викторович, канд. физ.-мат. наук, заведующий лабораторией анализа информационных ресурсов Научно-исследовательского вычислительного центра МГУ имени М.В. Ломоносова.