Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»
Направление «Интеллектуальный анализ данных»

Соболевский Федор Александрович
Б05-111

Применение больших языковых моделей для иерархической суммаризации текстов научных публикаций

Выпускная квалификационная работа бакалавра

Научный руководитель:
д. ф.-м. н. Воронцов Константин Вячеславович

Москва
2024

Содержание

Аннотация

В век экспоненциального роста количества доступной информации в мире особенно актуальной становится задача структурирования и систематизации научных знаний, а также повышения их доступности. Иерархическая организация основных идей и результатов в научных публикациях может позволить ускорить процесс получения читателем знаний и позволить ему двигаться при изучении темы от главного к деталям. Одним из видов структурированного представления текста являются интеллект-карты на основе предложений из текста. Поскольку человеческая обработка больших коллекций текстовых документов, особенно научных, занимает много времени и ресурсов, для решения задачи иерархической суммаризации необходимо разрабатывать автоматические методы, по качеству не уступающие ручной обработке.

Перспективным инструментом решения данной задачи являются большие языковые модели. В данной работе исследуется способность больших языковых моделей строить иерархические представления текстов научных публикаций на примере интеллект-карт на основе предложений. Поскольку для задачи автоматической иерархической суммаризации научных текстов на данный момент не существует достаточного количества обучающих данных и метрик для разностороннего и полного оценивания качества генерации, предварительно проводится работа по созданию новой коллекции иерархических сводок научных статей для обучения и тестирования моделей иерархической суммаризации и предлагаются новые способы оценивания результатов выполнения данной задачи.

Ключевые слова: *большие языковые модели, иерархическая суммаризация, интеллект-карты*

Введение

Цели работы.

- Формализовать задачу автоматического построения интеллект-карт по научным публикациям и предложить метрики оценивания таких карт, достаточно хорошо отражающие реальное качество генерации по структуре, связности и достоверности;
- Разработать и применить методику построения интеллект-карт по научным текстам экспертами с целью формирования выборки - золотого стандарта для обучения и валидации моделей автоматической иерархической суммаризации, а также с целью выработки четких требований и целей для автоматической генерации;
- Применить большие языковые модели (БЯМ) для генерации интеллект-карт по текстам научных статей и определить оптимальные запросы, позволяющие максимизировать качество генерации для каждой выбранной БЯМ;
- Проанализировать свойства генерируемых с помощью БЯМ карт, их достоинства и недостатки самих по себе и в сравнении со стандартами и определить границы применимости современных БЯМ для иерархической суммаризации научной литературы.

Обозначения и сокращения

- *БЯМ* — большая языковая модель (large language model);
- Под *картами знаний* или *интеллект-картами* в данной работе будут подразумеваться интеллект-карты на основе предложений (salient-sentence-based mind maps, SSM).
- *KSM* — интеллект-карта на основе главных выдержек (key-snippet based mind map).
- *ROUGE* - Recall-Oriented Understudy for Gisting Evaluation (статистическая метрика качества суммаризации).
- *BLEU* - Bilingual Evaluation Understudy (статистическая метрика качества генерации текста, основное применение — оценка машинного перевода).
- *ИАТ* — интеллектуальный анализ текста (text mining).

1 Постановка задачи

1.1 Задача иерархической суммаризации

Пусть дан документ (или коллекция документов) $\mathcal{D}$ — упорядоченный набор предложений, составленных из слов некоторого словаря V:

$$\mathcal{D} = (s_i)_{i=1}^{|\mathcal{D}|}, \quad \text{где } \forall i = 1, \dots |\mathcal{D}| \quad s_i = (w_{ij})_{j=1}^{l_i}, \quad w_{ij} \in V,$$

а также референсная карта $\mathcal{M}^* = (\mathcal{S}^*, E^*)$ и метрика качества $\mathcal{I} : (\mathcal{M}, \mathcal{D}, \mathcal{M}^*) \to \mathbb{R}$. Тогда требуется найти отображение $f^* : \mathcal{D} \to \mathcal{M} = (\mathcal{S}, E)$, максимизирующее данную метрику качества $\mathcal{I}$, где $\mathcal{M}$ — древовидная **иерархическая карта** (*интеллект-карта*, *карта знаний*), $\mathcal{S}$ — набор предложений, являющихся вершинами $\mathcal{M}$ и составленных из слов словаря V, $E \in \mathcal{S}^2$ — направленные иерархические связи между предложениями из $\mathcal{S}$, то есть ребра направленного графа $\mathcal{M}$:

$$f^* = \arg\max_f \mathcal{I}(f(\mathcal{D}), \mathcal{D}, \mathcal{M}^*).$$

1.2 Задача разработки критериев оценки карт знаний

Поскольку на данный момент нет четких критериев для автоматической оценки качества иерархической суммаризации в виде карт знаний, перед тестированием БЯМ необходимо разработать систему таких критериев. Пусть мы имеем сгенерированную моделью по документу $\mathcal{D}$ карту знаний $\mathcal{M}$ и карту-стандарт $\mathcal{M}^*$ по тому же документу, созданную экспертами. Требуется определить критерии

$$\mathcal{I}_k : (\mathcal{M}, \mathcal{D}, \mathcal{M}^*) \to \mathbb{R}$$

для автоматического оценивания интеллект-карт, отражающие интересующие нас аспекты качества генерации иерархического представления $\mathcal{M}$ относительно исходного документа $\mathcal{D}$ и карты-стандарта $\mathcal{M}^*$. Мера качества критерия — коэффициент корреляции с экспертной метрикой $\mathcal{I}_k^*$ оценки соответствующего аспекта качества по некоторой выборке карт $\mathbf{X}$.

Выделим пять основных реальных аспектов качества карты знаний:

- *соответствие цели*, для которой создаётся данное представление документа;
- *полнота* карты относительно документа, то есть содержание в ней всей необходимой пользователю информации на любом требуемом уровне абстракции и отсутствие лишнего;
- *непротиворечивость* иерархического представления как на уровне предложений, так и между ними;
- *связность* и *неизбыточность* карты как набора осмысленных предложений и их последовательностей (подразумевается, что любой путь из корня дерева произвольной длины представляет из себя связный, неизбыточный текст);
- *логичность* связей между вершинами в карте: степень логической связи между соединенными ребрами предложениями и корректность иерархии с логической точки зрения.

В предыдущей постановке задачи неизбежно возникает следующая проблема: количество экспертных карт и скорость их создания сильно ограничены, поэтому после отработки сравнительных критериев следует разработать способ оценивать карты без стандартов.

Пусть мы имеем только документ $\mathcal{D}$ и сгенерированную по нему моделью карту знаний $\mathcal{M}$. Требуется определить собственные критерии качества карты

$$\mathcal{I}_k : (\mathcal{M}, \mathcal{D}) \to \mathbb{R},$$

отражающие качество генерации иерархического представления $\mathcal{M}$ исходного документа $\mathcal{D}$ самого по себе, без сравнения с другими картами. Мерой качества автоматического критерия, помимо степени скоррелированности с экспертными оценками, может послужить также результат оценивания с помощью него экспертных карт $\mathcal{M}^*$, принимаемых за стандарт.

1.3 Задача оптимизации промптинга БЯМ

Основной инструмент оптимизации работы готовой БЯМ без её дополнительного дообучения — подбор оптимального текстового запроса для решения задачи (*промптинг*). Поскольку на данный момент нет достаточно быстрых алгоритмов поиска по всему множеству возможных запросов, а полный перебор запросов для каждой модели занимает слишком большое время, отдельной задачей в нашей работе является оптимизация промптинга БЯМ для цели генерации интеллект-карт и в целом.

Пусть дана выборка документов $\mathbf{X}$, языковая модель f, карта-стандарт $\mathcal{M}^*$ для заданной цели создания карты и некоторая метрика качества $\mathcal{I}$. Пусть также задано множество возможных запросов $\mathcal{Q}$, таких что вывод модели для каждого $(\mathcal{D}, Q) \in \mathbf{X} \times \mathcal{Q}$ соответствует требуемому формату, и содержащих формулировку цели, *общую для модели и экспертов* - создателей $\mathcal{M}^*$. Требуется найти оптимальный запрос $Q^* \in \mathcal{Q}$, такой что вывод модели при входе $(\mathcal{D}, Q^*)$ максимизирует метрику $\mathcal{I}$ по выборке $\mathbf{X}$:

$$Q^* = \arg\max_{Q \in \mathcal{Q}} \frac{1}{|\mathbf{X}|} \sum_{\mathcal{D} \in \mathbf{X}} \mathcal{I}(f(\mathcal{D}, Q), \mathcal{D},).$$

Для эффективного поиска оптимального запроса (*промптинга*) требуется также задать полное и неизбыточное множество запросов $\mathcal{Q}$ и эффективную *стратегию поиска оптимального запроса* в $\mathcal{Q}$.

2 Обзор

2.1 Методы суммаризации

Суммаризация текстов и БЯМ. Задача суммаризации текста представляет из себя задачу получения краткого представления $\mathcal{S}$ документа (или коллекции документов) $\mathcal{D}$. Выделяют два основных вида суммаризации: *экстрактивную* (extractive), использующую предложения исходного документа ($\mathcal{S} \subset \mathcal{D}$), и *абстрактивную*, то есть генерацию новых предложений на основе исходного текста ($\mathcal{S} \not\subset \mathcal{D}$). Также отдельно упоминается так называемая *гибридная суммаризация* (hybrid summarization), подразумевающая извлечение из документа важных предложений с последующим их преобразованием.

Хотя задача суммаризации появилась в научной литературе ещё во второй половине XX века [1], основной объем работ по суммаризации текстов был опубликован уже в XXI году, причем до начала бурного развития архитектур глубокого обучения основные методы построения сводок документов и их оценки были в большинстве своем экстрактивными, основанными на статистических приемах обработки текстов [2]. Абстрактивная суммаризация начала активно развиваться с момента появления трансформерных архитектур и других архитектур глубокой машинной обработки текста [3].

Особенных успехов в области удалось добиться с появлением больших языковых моделей, которые стали основным инструментом ля суммаризации текстов, так как показали значения метрик качества, которых до этого момента не удавалось добиться [4]. Более того, способности БЯМ к пониманию, обработке и генерации текста нашли свое применение не только для автоматической генерации сводок, но и для их же оценивания и корректировки [5]. Сравнение результатов работы БЯМ и человека по классической (линейной) суммаризации текстов на данный момент позволяет утверждать, что для некоторых типов текста машинная суммаризация с помощью БЯМ уже достигла уровня человека [6]. Авторы данной работы, однако, подчеркивают, что проблема оценки качества нейросетевой суммаризации и генерации текста в целом до сих пор остается открытой, поэтому нельзя утверждать о полной достаточности БЯМ для суммаризации. Применимость БЯМ для других видов суммаризации текстов все еще остается мало исследованной.

Иерархическая суммаризация. Идея структурированной суммаризации текстов как более эффективного способа суммаризации больших документов появилась в научной литературе еще в конце 2000-х гг. [7], однако впервые она была формализована в работе [8]. В первоначальной постановке задача иерархической суммаризации текста представляет из себя задачу генерации *иерархии из сводок* по исходной коллекции документов, в которой дочерние сводки раскрывают содержание элементов (например, предложений) более общих сводок. Метод, представленный в [8], подразумевает иерархическую кластеризацию предложений текста с последующей суммаризацией каждого кластера. Целевая функция в [8] отражает значимость выделенных предложений, неизбыточность и связность (как внутри элементов иерархии, так и между ними) полученной иерархической сводки. Хотя авторам удалось показать, что данный подход к суммаризации новостных текстов более предпочтителен среди пользователей, чем классические подходы к суммаризации новостных текстов,

этот подход не получил дальнейшего развития. Более применимыми стали подходы, основанные на генерации на основе текстов *интеллект-карт*.

Интеллект-карты. На сегодняшний день существует множество различных видов организации знаний в виде графовых структур: онтологии, карты концепций (concept maps) и другие, но в данной работе нас будут интересовать *интеллект-карты*, или *карты знаний*. В работах по автоматической генерации интеллект-карт выделяют два основных вида интеллект-карт: *интеллект-карты на основе значимых предложений* (salient sentence-based mind maps, SSM) и *интеллект-карты на основе главных выдержек* (key snippet-based mind maps, KSM). В данной работе нас будут интересовать именно SSM как форма иерархической организации *фактов* (в роли которых будут выступать связные предложения), однако следует подчеркнуть, что практическая значимость интеллект-карт в образовании и других областях исследовалась больше на примере KSM. Также обратим внимание на то, что интеллект-карты при применении их человеком зачастую не ограничиваются лишь иерархиями из текста и могут содержать как более сложные структурные элементы, так и нетекстовые элементы (например, визуальные).

После введения в употребление термина «mind map» Тони Бьюзеном в 1974 году интеллект-карты были экстенсивно изучены как инструмент представления, обработки и систематизации знаний. Многочисленные исследования применения интеллект-карт в школьном и высшем образовании как для презентации информации ученикам/студентам, так и для систематизации полученных знаний им же показали, что такой способ представления информации может заметно улучшить качество восприятия, запоминания и систематизации знаний студентами, в том числе при изучении научной литературы [9]. Подробный современный обзор применения интеллект-карт в образовании в XXI веке можно найти, например, в [10].

Автоматическая генерация интеллект-карт. В последнее десятилетие появился ряд работ по автоматической суммаризации текстов в виде интеллект-карт разных видов при помощи методов машинного обучения. Стоит отметить, что до этого были работы по генерации карт знаний/онтологий по текстам методами ИАТ, но в этих работах фокус больше направлен на моделирование взаимосвязей между отдельными словами/понятиями в тексте, чем между предложениями/фактами, поэтому для нашего исследования данные работы неактуальны.

В работе [11] был предложен метод интеллект-карт (как KSM, так и SSM) по текстам следующим способом: а) с помощью сравнения эмбеддингов предложений строится граф взаимосвязей между ними; б) по графу взаимосвязей строится интеллект-карта нужного вида. Данная идея нашла свое развитие в работе [12], в которой авторы предложили более эффективный способ превращения графа отношений между предложениями в интеллект-карту с помощью модуля дистилляции графа (graph refinement module). В работе [13] эта идея была усовершенствована применением вместо графа отношений между предложениями, строящегося по эмбеддингам предложений, графа соотнесенности (coreference graph, discourse graph), строящегося по принципу, описанному в работе [14]. В результате в [13] были полученные самые высокие значения используемых метрик качества, поэтому мы можем использовать данную модель в качестве базовой для сравнения результатов, полученных с помощью БЯМ, и результатов, достижимых с помощью специализированных

нейросетевых архитектур.

В недавнее время были начаты исследования способности БЯМ к генерации подобных интеллект-карт. В работе [15] с помощью промптинга больших языковых моделей строятся так называемые *StructSum* — структурированные сводки текстов для поиска конкретной информации, в частности, таблицы и интеллект-карты (KSM). Авторы применяют *самокритики* модели (critics) для улучшения качества генерации, такие как запросы по оцениванию самой моделью различных аспектов качества сгенерированного ею же StructSum и генерация вопросно-ответных пар по исходному тексту для проверки возможности находить нужную информацию из текста в полученной карте. Хотя тот формат карт и решаемые ими задачи, что исследуется в работе [15], несколько отличается от рассматриваемого в нашей работе, данное исследование подкрепляет предположение о том, что при достаточно изящной стратегии промптинга БЯМ могут стать качественным решением поставленной нами задачи.

2.2 Оценка качества суммаризации

Статистические критерии. Общепринятым подходом к оцениванию суммаризации с начала развития методов решения данной задачи является использование статистических критериев качества. Самые часто используемые из них, семейство метрик ROUGE [16], основаны на количестве совпадающих текстовых единиц, таких как n-граммы, последовательности слов и пары слов, между сгенерированной сводкой и экспертным резюме. Другие подобные метрики качества — BLEU [17], METEOR [18], MoverScore [19] и другие — схожи с ROUGE по принципу работы в том смысле, что они оценивают сходство стандартной и сгенерированной сводок на уровне слов, словосочетаний, n-грамм и других небольших семантических единиц.

Основной проблемой вышеперечисленных критериев является низкая репрезентативность статистического подхода в задаче оценки осмысленности, фактичности, связности и других более тонких аспектов сгенерированных сводок. Например, в работе [20] по результатам масштабного сравнительного исследования автоматических критериев качества и экспертных оценок искусственно сгенерированных сводок новостных текстов был сделан вывод о том, что экспертные оценки некоторых аспектов реального качества суммаризации, какие как связность и актуальность сводки, достаточно низко коррелируют со значениями автоматических метрик, что указывает на серьезную проблему с автоматическим оцениванием генерации текста статистическими критериями. Это указывает на необходимость использования более сложных критериев качества, учитывающих смысловую структуру исходного текста и его стандартных и искусственно сгенерированных сводок.

Критерии, основанные на БЯМ. Другим подходом к оцениванию качества суммаризации, ставшим довольно распространенным в последние пять лет, является оценивание суммаризации с помощью БЯМ. Из метрик, основанных на таком подходе, можно выделить BertScore [21], Shepherd [22] и Booookscore [23]. Несколько недавних работ также исследуют способность современных моделей по типу GPT оценивать качество суммаризации и искать ошибки в сгенерированных текстах. На данный момент такие методы также не являются полноценным решением проблемы оценивания качества суммаризации в силу неидеальности самих моделей, но они показывают многообещающие результаты. Более подробный обзор этих и других ме-

тодов оценивания суммаризации с помощью БЯМ можно найти в [4].

Отсутствие достаточно информативных метрик качества суммаризации остается основной проблемой в области суммаризации на сегодня. Во многих современных работах по автоматической суммаризации до сих пор используются простые статистические метрики по типу ROUGE и BLEU, причем зачастую смысл этих метрик не раскрывается, что делает сложным оценку реального качества генерируемых сводок. Хотя некоторыми исследователями были предприняты попытки систематического переосмысления оценивания качества суммаризации [20], [24], на сегодняшний день задача разработки достаточных метрик для полного, разностороннего автоматического оценивания качества суммаризации остается нерешенной.

3 Предлагаемый метод

Сбор экспериментальных данных. Для проверки работы БЯМ над задачей генерации интеллект-карт по научным публикациям необходима выборка таких карт, созданных людьми с достаточными компетенциями в данной задаче. Для этого и для отработки методики и целей построения иерархических сводок по научным документам первоначальна необходимо проведение самостоятельной работы по построению иерархических сводок научных статей. Предполагается создание набора первоначальных карт, в ходе построения которых, во-первых, должны быть установлены принципы, которыми следует руководствоваться при создании карты знаний, и, во-вторых, должны быть определены цели создания таких карт. Сформулируем несколько примерных целей создания карты знаний по научной публикации — получение минимальных знаний, необходимых для:

- воспроизведения результатов авторов статьи студентом или младшим научным сотрудником, разбирающимся в предметной области на базовом уровне;
- выделения наиболее важного, нового, значимого результата для его популяризации или включения в образовательный курс по предметной области данного исследования;
- выделения основного результата для упоминания в научном обзоре по предметной области;
- подготовки пересказа статьи на научном семинаре, максимально близкого к тому, что могли бы рассказать сами авторы, желая донести свои результаты до профессионального сообщества в своей предметной области.

Вполне естественно, что данными целями спектр применимости карт знаний не ограничивается, но для определенности в исследовании мы зафиксируем именно их.

Агрегация экспертных мнений. Так как мнения экспертов по поводу оптимальных способов построения карт знаний могут розниться, эти мнения нужно агрегировать. Мы предлагаем следующую методику объединения экспертных усилий для совместного создания интеллект-карт по научным текстам: проведение научных семинаров в формате обсуждения интеллект-карт по статьям с последующим построением общей карты. Пусть мы собрали N экспертов и задали K целей создания интеллект-карты, тогда на выходе мы имеем по каждой обработанной статье $N +$ карт, K из которых выбираются нами в качестве стандартных для дальнейших исследований.

Отбор критериев качества. Вышеупомянутые научные семинары можно использовать не только с целью создания стандартных интеллект-карт, но и для сбора экспертных оценок аспектов качества человеческих и сгенерированных искусственно интеллект-карт по рассматриваемым статьям. Это позволит вместе с созданием выборки собрать также данные для корреляционного анализа экспертных оценок и значений возможных критериев качества интеллект-карт, необходимого для последующего выбора критериев для оценивания машинной генерации иерархических карт автоматически.

Список литературы

[1] Luhn Hans Peter. The automatic creation of literature abstracts // IBM Journal of research and development. — 1958. — Vol. 2, no. 2. — P. 159–165.

[2] Text summarization techniques: a brief survey / Allahyari Mehdi, Pouriyeh Seyedamin, Assefi Mehdi, Safaei Saeid, Trippe Elizabeth D, Gutierrez Juan B, and Kochut Krys // arXiv preprint arXiv:1707.02268. — 2017.

[3] Automatic text summarization: A comprehensive survey / El-Kassas Wafaa S, Salama Cherif R, Rafea Ahmed A, and Mohamed Hoda K // Expert systems with applications. — 2021. — Vol. 165. — P. 113679.

[4] A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods / Jin Hanlei, Zhang Yang, Meng Dan, Wang Jun, and Tan Jinghua // arXiv preprint arXiv:2403.02901. — 2024.

[5] Large language models are diverse role-players for summarization evaluation / Wu Ning, Gong Ming, Shou Linjun, Liang Shining, and Jiang Daxin // CCF International Conference on Natural Language Processing and Chinese Computing / Springer. — 2023. — P. 695–707.

[6] Pu Xiao, Gao Mingqi, Wan Xiaojun. Summarization is (almost) dead // arXiv preprint arXiv:2309.09558. — 2023.

[7] Yang Christopher C, Wang Fu Lee. Hierarchical summarization of large documents // Journal of the American Society for Information Science and Technology. — 2008. — Vol. 59, no. 6. — P. 887–902.

[8] Hierarchical summarization: Scaling up multi-document summarization / Christensen Janara, Soderland Stephen, Bansal Gagan, et al. // Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). — 2014. — P. 902–912.

[9] Guerrero Jose M, Ramos Pilar. Mind mapping for reading and understanding scientific literature // International Journal of Current Advanced Research. — 2015. — Vol. 4, no. 11. — P. 485–487.

[10] From Tradition to Innovation: Mind Map Generation in Higher Education / Mitra Aditya Rama, Samosir Feliks Victor Parningotan, Hudi Robertus, and Tarigan Riswan Effendi // Ultima InfoSys: Jurnal Ilmu Sistem Informasi. — 2023. — Vol. 14, no. 2. — P. 71–78.

[11] Revealing Semantic Structures of Texts: Multi-grained Framework for Automatic Mind-map Generation. / Wei Yang, Guo Honglei, Wei Jin-Mao, and Su Zhong // IJCAI. — 2019. — P. 5247–5254.

[12] Efficient Mind-Map generation via Sequence-to-Graph and reinforced graph refinement / Hu Mengting, Guo Honglei, Zhao Shiwan, Gao Hang, and Su Zhong // arXiv preprint arXiv:2109.02457. — 2021.

[13] Coreference Graph Guidance for Mind-Map Generation / Zhang Zhuowei, Hu Mengting, Bai Yinhao, and Zhang Zhen // Proceedings of the AAAI Conference on Artificial Intelligence. — 2024. — Vol. 38. — P. 19623–19631.

[14] Discourse-aware neural extractive text summarization / Xu Jiacheng, Gan Zhe, Cheng Yu, and Liu Jingjing // arXiv preprint arXiv:1910.14142. — 2019.

[15] Jain Parag, Marzoca Andreea, Piccinno Francesco. Structsum Generation for Faster Text Comprehension // arXiv preprint arXiv:2401.06837. — 2024.

[16] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries // Text summarization branches out. — 2004. — P. 74–81.

[17] Bleu: a method for automatic evaluation of machine translation / Papineni Kishore, Roukos Salim, Ward Todd, and Zhu Wei-Jing // Proceedings of the 40th annual meeting of the Association for Computational Linguistics. — 2002. — P. 311–318.

[18] Banerjee Satanjeev, Lavie Alon. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. — 2005. — P. 65–72.

[19] MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance / Zhao Wei, Peyrard Maxime, Liu Fei, Gao Yang, Meyer Christian M, and Eger Steffen // arXiv preprint arXiv:1909.02622. — 2019.

[20] Summeval: Re-evaluating summarization evaluation / Fabbri Alexander R, Kryściński Wojciech, McCann Bryan, Xiong Caiming, Socher Richard, and Radev Dragomir // Transactions of the Association for Computational Linguistics. — 2021. — Vol. 9. — P. 391–409.

[21] Bertscore: Evaluating text generation with bert / Zhang Tianyi, Kishore Varsha, Wu Felix, Weinberger Kilian Q, and Artzi Yoav // arXiv preprint arXiv:1904.09675. — 2019.

[22] Shepherd: A critic for language model generation / Wang Tianlu, Yu Ping, Tan Xiaoqing Ellen, O'Brien Sean, Pasunuru Ramakanth, Dwivedi-Yu Jane, Golovneva Olga, Zettlemoyer Luke, Fazel-Zarandi Maryam, and Celikyilmaz Asli // arXiv preprint arXiv:2308.04592. — 2023.

[23] Booookscore: A systematic exploration of book-length summarization in the era of llms / Chang Yapei, Lo Kyle, Goyal Tanya, and Iyyer Mohit // arXiv preprint arXiv:2310.00785. — 2023.

[24] Benchmarking large language models for news summarization / Zhang Tianyi, Ladhak Faisal, Durmus Esin, Liang Percy, McKeown Kathleen, and Hashimoto Tatsunori B // Transactions of the Association for Computational Linguistics. — 2024. — Vol. 12. — P. 39–57.