Аннотация

В век экспоненциального роста количества доступной информации в мире особенно актуальной становится задача структурирования и систематизации научных знаний, а также повышения их доступности. Структурированная организация основных идей и результатов из научной публикации может позволить ускорить процесс получения читателем знаний из нее. Одним из видов структурированного представления текста, позволяющих двигаться при изучении темы от главного к деталям, являются иерархические сводки. Поскольку человеческая обработка научных текстов с целью создания иерархической сводки занимает много времени, возникает необходимость разрабатывать автоматические методы иерархической суммаризации, по качеству не уступающие ручной суммаризации.

Перспективным инструментом решения данной задачи являются большие языковые модели (БЯМ). В данной работе исследуется способность больших языковых моделей строить иерархические представления текстов научных публикаций. Основным методом оценки качества иерархической суммаризации, как и обычной суммаризации, является оценка сходства с эталонной сводкой, созданной экспертом. Поскольку на данный момент не существует выборок для задачи иерархической суммаризации научных текстов, предварительно создается выборка иерархических сводок по ряду научных статей для сравнения со сгенерированными автоматически. Иерархическая суммаризация с помощью БЯМ оценивается в сравнении со сводками из этой выборки с учетом различных аспектов сходства иерархических сводок, таких как структура и семантика сводки.

Применяемые до сих пор методы сравнения текстовых иерархий основаны на сравнении их на лексическом уровне и, как показано в данной работе, слабо учитывают их структуру и семантику по отношению к фразировке. В связи с этим в данной работе предлагается также новый метод сравнения текстовых деревьев — расстояние редактирования текстовых деревьев (TTED), основанное на расстоянии редактирования и оценке семантической близости с помощью языковых моделей. Для оценки информативности функции расстояния между текстовыми деревьями как агрегации разных аспектов их различия вводятся R_S- и R_M-коэффициенты, отражающие чувствительность функции сходства к семантическим и структурным различиям текстовых деревьев по отношению к перефразированию текстов в вершинах, а также предлагаются несмещенные оценки на эти коэффициенты по случайной выборке текстовых деревьев. С помощью этих оценок дается количественная оценка качества предложенной метрики и ее модификаций в сравнении с базовой, использованной в предыдущих работах по теме.

Ключевые слова: *большие языковые модели, иерархическая суммаризация, интеллект-карты, текстовые деревья, расстояние редактирования*

Содержание

Введение

Актуальность темы. С развитием средств хранения, обработки и передачи информации и накоплением данных человечеством количество информации в мире экспоненциально растет, и научные знания не являются исключением. С увеличением скорости появления новых научных публикаций и накоплением старых материалов возрастает необходимость как в быстром поиске публикаций, актуальных для пользователя, так и в эффективном извлечении знаний из них. При необходимости обработки больших объемов научных текстов у исследователя возникает потребность в эффективной организации информации, позволяющей вникать в содержание научных исследований с нужной степенью детализации, двигаясь при изучении темы от главного к деталям.

Методом удовлетворения данной потребности представляется *иерархическая суммаризация* научных публикаций, то есть представление их в виде иерархической структуры, в которой на верхних уровнях иерархии находятся ключевые аспекты исследования, а каждая дочерняя вершина детализирует родительскую. Такое представление знаний информации из научного труда потенциально может позволить изучить его с нужным уровнем детализации по каждому отдельному аспекту исследования и извлечь ровно тот объем информации из него, который необходим читателю. Такое представление знаний также может стать способом повысить доступность научных знаний по теме, с помощью которого можно кратко и доступно разобраться в теме и затем при необходимости изучить интересующие детали.

Представление знаний в виде иерархических структур не является новой идеей. В литературе давно известен термин «интеллект-карта» (mind map), обозначающий древовидную карту, иерархически раскрывающую тему от главных понятий к деталям [1]. Интеллект-карты показали себя как инструмент, позволяющий улучшить восприятие и запоминание информации [2], однако самостоятельное построение интеллект-карты по тексту научной публикации является достаточно затратным по времени занятием. Автоматическая генерация интеллект-карт по научным статьям представляется способом устранить эту проблему и предоставить удобную для восприятия организацию знаний читателю без дополнительных затрат времени.

В последние годы одним из самых универсальных инструментов для генерации и обработки текста стали большие языковые модели (БЯМ). Выдающиеся способности таких моделей к генерации связного текста уже нашли свое применение во многих задачах обработки текстов, в том числе в задаче суммаризации, причем современные БЯМ уже способны показывать в данной задаче результаты, по качеству сопоставимые с человеческими [3]. Несмотря на это, хотя в последние годы появилось небольшое число работ, изучающих способность БЯМ генерировать структурированные представления информации [4], исследование БЯМ в приложении к задаче иерархической суммаризации еще только предстоит провести.

Немаловажной частью данного исследования является разработка нового подхода к оцениванию иерархической суммаризации и создание новой выборки для этой цели. Несмотря на наличие ряда работ по теме [4–10], иерархическая суммаризация является относительно малоизученной задачей, и на данный момент не существует общепринятых методов оценивания автоматической генерации иерархических сводок, а выборки для этой задачи ограничиваются областью новостных текстов и экстрактивным подходом к иерархической суммаризации. Используемый в последних

работах по теме метод сравнения иерархических сводок с экспертными основан, как показано в данном исследовании, слабо отражает значимые различия между иерархическими сводками по сравнению с различиями в формулировках. Это обосновывает необходимость разработки новой метрики сходства текстовых иерархий и нового подхода к оцениванию подобных метрик сходства.

Цели работы.

- Формализовать задачу автоматической иерархической суммаризации как задачу многокритериальной оптимизации.
- Сформировать выборку для обучения и валидации моделей автоматической иерархической суммаризации.
- Применить БЯМ для иерархической суммаризации текстов научных статей и определить оптимальный метод работы с моделью, позволяющий максимизировать качество генерации для выбранной БЯМ.
- Формализовать требования к адекватности метрики на множестве текстовых иерархий как метрики на множестве объектов, обладающих различными аспектами сходства.
- Предложить новую агрегированную метрику качества для задачи иерархической суммаризации, отражающую прежде всего значимые различия текстовых иерархий.
- Проанализировать генерируемые с помощью БЯМ иерархические сводки в сравнении с созданными человеком и определить границы применимости современных БЯМ для иерархической суммаризации научной литературы.

Научная новизна. Задача иерархической суммаризации на данный момент является достаточно малоизученной — имеющиеся исследования по данной теме, проведенные в разные моменты времени за последнее десятилетие, рассматривают данную задачу по-разному и по-разному оценивают качество иерархической суммаризации, вследствие чего в литературе не сложилось единообразного взгляда на данную задачу и оценивание качества в ней [4–10]. В данной работе предлагается общая постановка задача иерархической суммаризации, обобщающая имеющиеся.

Существующие на данный момент подходы к оцениванию иерархической суммаризации либо не являются автоматическими (например, оценивание с помощью краудсорсинга или экспертного труда, как в работах [4, 5]), либо, как показано в данном исследовании, недостаточно информативно отражают реальные различия иерархических сводок. В данном исследовании предлагается новый способ сравнения текстовых деревьев как более информативный способ оценивания иерархической суммаризации. Более того, в данной работе предлагается способ оценивания качества самих метрик на множестве текстовых деревьев, позволяющий исследовать информативность как существующих метрик, так и тех, которые, возможно, появятся в будущем.

Насколько автору известно, исследований на тему применения БЯМ к иерархической суммаризации или выборок для иерархической суммаризации текстов научных публикаций на текущий момент нет, поэтому работа также позиционируется как отправная точка для дальнейших исследований по теме, предлагая небольшую выборку для оценивания иерархической суммаризации научных текстов и качественный анализ качества иерархической суммаризации таких текстов с помощью БЯМ.

Теоретическая значимость. Иерархическая суммаризация является потенциально инновационным методом представления информации, однако данная задача до сих пор не поставлена в едином формальном виде. Данное исследование предлагает обобщенную формальную постановку задачи в качестве теоретической основы для дальнейших исследований по теме вместе с метрикой для оценки качества решения задачи. Более того, предложенные в данной работе коэффициенты качества метрик на множестве текстовых деревьев не только формализуют требования к информативности подобных метрик — они также позволяют оценить качество других метрик, которые могут быть предложены в данной задаче, в сравнении друг с другом и с существующими.

Практическая значимость. В данной работе, прежде всего, предлагаются методы построения иерархических сводок текстов с помощью БЯМ, которые можно применить и развить как в дальнейших исследованиях по теме, так и в личном пользовании при взаимодействии с разговорными БЯМ. Также автором предлагаются выборки для задач оценивания качества метрик для сравнения текстовых деревьев и иерархической суммаризации научных текстов, которые можно как использовать для будущих исследований по теме, так и расширить и дополнить по методологиям, описанным в данной работе. Методы иерархической суммаризации и оценки ее качества потенциально можно внедрить, например, в поисково-рекомендательные системы по научной литературе.

В данной работе предложена новая метрика на множестве текстовых деревьев, TTED, показавшая значительно лучшие результаты в терминах информативности, чем используемые до этого методы. Эту метрику можно применить как для оценивания иерархической суммаризации, так и для любых других задач, где возникает потребность в сравнении иерархий из текста. Для экспериментов в рамках данного исследования метрика TTED была реализована на языке Python и предоставлена в открытом доступе для дальнейшего использования.

Степень достоверности и апробация работы. Результаты работы по разработке метрики TTED на множестве текстовых деревьев были представлены на 67-й Всероссийской научной конференции МФТИ в докладе «Метод оценки сходства текстовых деревьев с помощью расстояния редактирования и языковых моделей» в Секции проблем интеллектуального анализа данных, распознавания и прогнозирования. Весь код, использованный в данной работе для проведения вычислительных экспериментов, находится в открытом доступе для репликации полученных результатов по ссылке: `https://github.com/intsystems/Sobolevsky-BS-Thesis`.

Обозначения и сокращения

- *БЯМ* — большая языковая модель (large language model).
- *ИАТ* — интеллектуальный анализ текста (text mining).
- *ИИ* — искусственный интеллект.
- *BLEU* — Bilingual Evaluation Understudy [11].
- *KSM* — интеллект-карта на основе главных выдержек (key-snippet based mind map [7]).
- *SSM* — интеллект-карта на основе значимых предложений (salient-sentence-based mind map [7]).
- *NLP* — обработка естественного языка (natural language processing).
- *ROUGE* — Recall-Oriented Understudy for Gisting Evaluation [12].
- *TED* — расстояние редактирования дерева (tree edit distance [13]).
- *TTED* — расстояние редактирования текстового дерева (text tree edit distance).

1 Постановка задачи

1.1 Текстовые иерархии как объект исследования

Объектом данного исследования являются иерархические сводки текстовых документов. Иерархическая сводка по своей структуре есть ничто иное, как текстовое дерево, то есть дерево, в вершинах которого находятся фрагменты текста. Определим этот объект формально. Пусть задан словарь $\mathcal{W}$ и соответствующее множество $\mathcal{S}$ текстов над данным словарем:

$$\forall s \in \mathcal{S} \quad s = (w_j)_{j=1}^{|s|}, \quad w_j \in \mathcal{W}.$$

Определим текстовое дерево $T = (V, E)$, $E \subset V^2$, для каждой вершины $v \in V$ которого задан текст $s(v) \in \mathcal{S}$. Обозначим множество рассматриваемых текстовых деревьев как $\mathcal{T}$. Текстовые деревья $T \in \mathcal{T}$ будут объектом генерации и сравнения в данном исследовании.

1.2 Задача иерархической суммаризации

Пусть дан документ (или коллекция документов) $\mathcal{D}$ — упорядоченный набор фрагментов текста из $\mathcal{S}$: $\mathcal{D} = (s_i)_{i=1}^{|\mathcal{D}|}$, $s_i \in \mathcal{S}$ и задана цель иерархической суммаризации в виде текстового запроса $q \in \mathcal{S}$, а также эталонная иерархическая сводка $T^* \in \mathcal{T}$ данного документа, построенная экспертом с заданной целью. Пусть также задана метрика качества $\mathcal{I}$ генерации текстовых деревьев $T \in \mathcal{T}$ по документам, в общем случае зависящая от экспертной сводки T^*, цели q и самого документа $\mathcal{D}$, то есть $\mathcal{I} : (T, T^*, q, \mathcal{D}) \mapsto x \in \mathbb{R}$. Тогда требуется найти отображение $f : \mathcal{D} \mapsto T \in \mathcal{T}$, максимизирующее данную метрику качества $\mathcal{I}$:

$$\mathcal{I}(f(\mathcal{D}), T^*, q, \mathcal{D}) \longrightarrow \max_f. \tag{1}$$

В более общем случае можно определить набор метрик $\mathcal{I}_k$ такого вида, отражающих различные аспекты качества генерации иерархической сводки T. Из основных аспектов качества иерархической суммаризации можно выделить следующие:

- Сходство с экспертной сводкой T^* по таким аспектам, как структура сводки, ее смысловое содержание, ранжирование фактов в сводке и др.
- Качество сводки как краткого представления документа $\mathcal{D}$, например, полнота сводки или степень согласованности T и $\mathcal{D}$.
- Соответствие поставленной цели иерархической суммаризации, например, в терминах фактического соответствия запросу q.
- Качество сводки самой по себе, например, отсутствие избыточности, связность текста внутри вершин и между вершинами, непротиворечивость, соответствие цели генерации и т. д.

Пусть задана функция $\mathcal{A}(\mathcal{I}_1, \ldots, \mathcal{I}_K) : \mathbb{R}^K \longrightarrow \mathbb{R}$, каким-то образом агрегирующая метрики аспектов качества $\mathcal{I}_k$. Тогда оптимизационную задачу (1) можно записать в более общем виде как

$$\mathcal{A}\left(\mathcal{I}_1(f(\mathcal{D}), T^*, q, \mathcal{D}), \ldots, \mathcal{I}_K(f(\mathcal{D}), T^*, q, \mathcal{D})\right) \longrightarrow \max_f. \tag{2}$$

1.3 Требования к метрике на множестве текстовых деревьев.

Чтобы определить адекватную метрику на множестве текстовых деревьев, сформулируем некоторые требования к произвольной метрике в пространстве таких объектов. Пусть задана функция семантической (смысловой) близости текстовых фрагментов: $r : \mathcal{S}^2 \to [0, +\infty)$. Для вершин $v, v' \in V$ дерева $T = (V, E)$ обозначим $r(v, v') := r(s(v), s(v'))$, а $r(v) := r(s(v), \lambda)$, где λ — пустая строка. Требуется определить функцию сходства $\rho : \mathcal{T}^2 \to [0, +\infty)$, удовлетворяющую следующим требованиям:

1. Симметричность: $\rho(T, T') = \rho(T', T)$.

2. Равенство нулю в случае равенства аргументов: $\rho(T, T) = 0$.

3. ρ удовлетворяет неравенству треугольника:

$$\forall T, T', T'' \in \mathcal{T} \quad \rho(T, T'') \leq \rho(T, T') + \rho(T', T''). \tag{3}$$

4. Существует некоторая неубывающая функция $f : [0, +\infty) \to [0, +\infty)$, такая что:

   (a) Если T' получено из T добавлением в T вершины v, то $\rho(T, T') = f(r(v))$;

   (b) Если T' получено из T удалением из T вершины v, то $\rho(T, T') = f(r(v))$;

   (c) Если T' получено из T заменой вершины v на v', то $\rho(T, T') = f(r(v, v'))$.

Обозначим теперь некоторые требования к существенности различий между текстовыми деревьями, отражаемых функцией их сходства. Во-первых, естественно требовать от метрики, чтобы она отражала различия текстовых деревьев как по своей *семантике* (то есть по смысловому содержанию), так и по *структуре*. Во-вторых, информативная метрика должна слабо реагировать на несущественные отличия — например, на *перефразирование* текстов в вершинах дерева. Следовательно, средние значения метрики для деревьев, различающихся по первым двум признакам, должны быть как можно меньше среднего значения расстояния между деревьями, полученными друг из друга перефразированием.

Формализуем эти требования. Пусть задано вероятностное пространство $(\mathcal{T}, 2^{\mathcal{T}}, \mathbb{P}_{\mathcal{T}})$, где $\mathbb{P}_{\mathcal{T}}$ — вероятностная мера, задающая распределение текстовых деревьев T из $\mathcal{T}$. Обозначим $P(T)$ как множество деревьев, которые можно получить из T перефразированием его меток, $S(T)$ — множество деревьев, составленных из того набора вершин с теми же текстами в них, что и T (но отличающихся по структуре), $M(T)$ — набор деревьев с такой же структурой, как у T, но с разной семантикой. Для конкретности определим последнее множество как множество $\mathcal{T}_{\sim T}$ деревьев с той же структурой, как у T, за вычетом самого дерева T и его парафразов: $M(T) = \mathcal{T}_{\sim T} \setminus (P(T) \cup \{T\})$. Тогда качественные требования, сформулированные выше, будут отражены формально в следующих соотношениях:

$$\mathbb{E}_{T' \sim P(T),\ T'' \sim S(T)} \left[\frac{\rho(T, T')}{\rho(T, T'')} \right] \ll 1, \quad \mathbb{E}_{T' \sim P(T),\ T'' \sim S(T)} \left[\frac{\rho(T, T')}{\rho(T, T'')} \right] \ll 1 \quad \forall T \in \mathcal{T}. \tag{4}$$

Здесь для краткости записью $T' \sim P(T)$ и аналогичными записями для остальных множеств обозначено семплирование текстовых деревьев из распределений над соответствующими множествами. Знаменатель данных выражений не обращается в нуль, коль скоро ρ является корректной метрикой. Обозначим

$$\mathbb{E}_{T' \sim P(T),\ T'' \sim S(T)} \left[\frac{\rho(T, T')}{\rho(T, T'')} \right] := r_S(\rho, T), \quad \mathbb{E}_{T' \sim P(T),\ T'' \sim S(T)} \left[\frac{\rho(T, T')}{\rho(T, T'')} \right] := r_M(\rho, T). \tag{5}$$

Введем коэффициенты информативности метрики ρ, не зависящие от выбранного дерева $T \in \mathcal{T}$:

$$R_S(\rho) = \mathbb{E}_{T \sim \mathcal{T}}[r_S(\rho, T)], \quad R_M(\rho) = \mathbb{E}_{T \sim \mathcal{T}}[r_M(\rho, T)] \tag{6}$$

Требования (4) можно записать в более простом виде: $R_S(\rho) \ll 1$, $R_M(\rho) \ll 1$. Тогда соответствующие задачи оптимизации, решаемые в данной работе, будут выглядеть следующим образом:

$$R_S(\rho) \longrightarrow \min_{\rho}, \quad R_M(\rho) \longrightarrow \min_{\rho}. \tag{7}$$

Следует заметить, что коэффициенты информативности, введенные таким образом, будут зависеть от распределений вероятностей на соответствующих множествах текстовых деревьев при выбранном методе семплирования, и для их точного вычисления пришлось бы не только перебрать все деревья из множества $\mathcal{T}$ и соответствующие им модификации, но и определить вид распределений над соответствующими множествами. Однако, как будет показано далее, можно получить несмещенную оценку на эти коэффициенты по случайной выборке, не зависящую от вида распределений вероятностей семплирования.

2 Обзор

2.1 Методы суммаризации

Суммаризация текстов и БЯМ. Задача суммаризации текста представляет собой задачу получения краткого представления $\mathcal{S}$ документа (или коллекции документов) $\mathcal{D}$. Выделяют два основных подхода к суммаризации: *извлекающий* (extractive), использующий предложения исходного документа ($\mathcal{S} \subset \mathcal{D}$), и *генерирующий* (abstractive), то есть генерацию новых предложений на основе исходного текста ($\mathcal{S} \not\subset \mathcal{D}$) [14]. Также отдельно упоминается так называемая *гибридная суммаризация* (hybrid summarization), подразумевающая извлечение из документа важных предложений с последующим их преобразованием [15].

Хотя задача суммаризации появилась в научной литературе ещё во второй половине XX века [16], основной объем работ по суммаризации текстов был опубликован уже в XXI веке, причем до начала бурного развития архитектур глубокого обучения основными были извлекающие методы построения сводок документов и их оценки, основанные на статистических приемах обработки текстов [17]. Генерирующая суммаризация начала активно развиваться с момента появления трансформерных архитектур и других архитектур глубокого обучения [15].

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 [18]. 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 [19]. 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 [3]. Авторы данной работы, однако, подчеркивают, что проблема оценки качества суммаризации и генерации текста в целом с помощью нейросетей до сих пор остается открытой, поэтому нельзя утверждать, что БЯМ для суммаризации. Применимость БЯМ для других видов суммаризации текстов, в том числе иерархической, 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. [20], однако впервые она была формализована в работе [5]. В первоначальной постановке задача иерархической суммаризации текста представляет собой задачу генерации *иерархии из сводок* по исходной коллекции документов, в которой дочерние сводки раскрывают содержание родительской. Метод, представленный в [5], подразумевает иерархическую кластеризацию предложений текста с последующей суммаризацией каждого кластера. Целевая функция в [5] отражает значимость выделенных предложений, избыточность 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 для пользователей, чем классические подходы, данная методика получила свое развитие лишь в небольшом количестве работ [9, 10]. Однако впоследствии появился несколько другой подход к структури-

рованной суммаризации, основанный на генерации *интеллект-карт* по текстам.

Интеллект-карты. На сегодняшний день существует множество различных видов организации знаний в виде графов: онтологии, карты концепций (concept maps) и другие, но в рамках данного исследования наиболее релевантными являются *интеллект-карты* [1], поскольку они являются частным случаем иерархических сводок как способа структурированного представления информации в порядке от главного к деталям. 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 [7]) и *интеллект-карты на основе главных выдержек* (key snippet-based mind maps, KSM [7]). Данной работе более актуальны именно SSM как форма иерархической организации связного текста, однако следует подчеркнуть, что практическая значимость интеллект-карт в образовании и других областях исследовалась больше на примере KSM. Также следует обратить внимание на то, что интеллект-карты при применении их человеком зачастую содержат не только текст, но и визуальные элементы.

С момента появления в литературе термина «mind map» [1] 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, так и для систематизации полученных знаний показали, что такой способ представления информации может заметно улучшить качество восприятия, запоминания и систематизации знаний студентами, в том числе при изучении научной литературы [2, 21, 22]. Подробный современный обзор применения интеллект-карт в образовании в XXI веке можно найти, например, в [23].

Автоматическая генерация интеллект-карт. За последние 6 лет появился ряд работ по автоматической суммаризации текстов в виде интеллект-карт разных видов при помощи методов машинного обучения. Стоит отметить, что до этого было проведено не одно исследование по генерации интеллект-карт/онтологий по текстам методами ИАТ [24, 25], но в этих работах целью исследования является моделирование взаимосвязей между отдельными словами/понятиями в тексте, а не между предложениями/фактами, поэтому в рамках данного исследования эти работы не совсем релевантны.

В работе [7] предложен метод построения интеллект-карт (KSM и SSM) по текстам следующим способом: а) с помощью сравнения *эмбеддингов*, то есть векторных представлений, предложений, полученных при помощи нейронной сети, строится граф взаимосвязей между ними; б) по графу взаимосвязей строится интеллект-карта нужного вида. Данная идея нашла свое развитие в работе [8], в которой авторы предложили более эффективный способ превращения графа взаимосвязей между предложениями в интеллект-карту с помощью модуля дистилляции графа (graph refinement module). В работе [6] эта идея была усовершенствована применением вместо графа взаимосвязей между предложениями, строящегося по эмбеддингам предложений, графа кореференций (coreference graph, discourse graph), строящегося по принципу, описанному в работе [26]. Таким образом, в [6] был представлен усовершенствованный метод извлекающей иерархической суммаризации текстов. Следует отметить, что данный метод, во-первых, реализует извлекающий тип суммаризации,

а, во-вторых, больше предназначен для работы с небольшими текстами (например, с новостными, которые и исследовались в работе), поскольку время работы этого метода для более крупных текстов (например, научных статей) очень велико, поэтому прямое сравнение данного метода с БЯМ в этой работе затруднительно.

В недавнее время были начаты исследования способности БЯМ к генерации подобных интеллект-карт. В работе [4] 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 *самокритику* модели (critics) для улучшения качества генерации в виде запросов по оцениванию самой моделью различных аспектов качества сгенерированного ею же StructSum и генерации вопросно-ответных пар по исходному тексту для проверки возможности находить нужную информацию из текста в полученной карте. Хотя формат сводок и постановка задачи в работе [4] несколько отличаются от рассматриваемых в данной, это исследование подкрепляет предположение о том, что при достаточно изящной стратегии промптинга БЯМ могут генерировать структурированные представления текстов и качественно решать задачу иерархической суммаризации.

Необходимо также упомянуть появление в последние годы многочисленных коммерческих сервисов для генерации интеллект-карт с помощью ИИ (конкретно, с помощью БЯМ и инструментов NLP). Среди компаний, предоставляющих подобные сервисы, можно отметить MyMap AI[1], MindMap AI[2], Monica[3], Mapify[4] и некоторые другие. Несмотря на относительную распространенность подобных сервисов, реальное качество иерархической суммаризации с помощью подобных сервисов в сравнении с человеческой остается неизмеренным. Это обуславливает необходимость в научном исследовании подобных методов для определения границ применимости БЯМ в задаче иерархической суммаризации.

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, метрики из семейства ROUGE [12], 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 [11], METEOR [27], MoverScore [28] и другие — схожи с ROUGE по принципу работы в том смысле, что они оценивают сходство стандартной и сгенерированной сводок на уровне лексических 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 [29] по результатам масштабного сравнительного исследования автоматических критериев качества и экспертных оценок искусственно сгенерированных сводок но-

[1] `https://www.mymap.ai/mindmap`

[2] `https://mindmapai.app/`

[3] `https://monica.im/tools/ai-mind-map-maker`

[4] `https://mapify.so/`

востных текстов был сделан вывод о том, что экспертные оценки таких аспектов реального качества суммаризации, как. например, 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 сводок.

Критерии, основанные на БЯМ. Другим подходом к оцениванию качества суммаризации, ставшим довольно распространенным в последние пять лет, является оценивание суммаризации с помощью БЯМ. Одной из довольно простых, но уже широко применяемых метрик сходства текстов, основанных на БЯМ, является BERTScore [30]. Суть метода заключается в сравнении текстов по близости их представлений (*эмбеддингов*), полученных с помощью трансформерной модели BERT [31]. Данный метод показал лучшие результаты в терминах корреляции с экспертными оценками суммаризации, чем ROUGE, в силу того, что моделирование текстов с помощью нейросетей позволяет улавливать их семантику лучше, чем статистические методы.

Ряд недавних работ также исследует способность современных БЯМ оценивать качество суммаризации и искать ошибки в сгенерированных текстах [19, 32, 33]. На данный момент такие методы также не являются полноценным решением проблемы оценивания качества суммаризации в силу неидеальности самих языковых моделей, но они показывают многообещающие результаты [18].

Отсутствие общепринятой информативной метрики качества суммаризации остается основной проблемой в области суммаризации на сегодня. Во многих современных работах по теме автоматической суммаризации текстов 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 [29], [34], на сегодняшний день задача разработки достаточных метрик для полного, разностороннего автоматического оценивания качества суммаризации остается нерешенной.

Определение семантического сходства. На сегодняшний день лучшие результаты при решении задач оценки семантической близости текстов и детектирования парафразов были достигнуты с использованием моделей на основе трансформерных архитектур; в частности, с использованием моделей на основе архитектуры BERT и им подобных. Например, в исследовании [35] по результатам применения различных методов детектирования перефразирования на нескольких выборках было установлено существенное преимущество нейросетевых моделей перед лексическими методами. В данном исследовании было также установлено, что лучшие результаты получаются при сравнении эмбеддингов, полученных с помощью нейросетей, при помощи косинусного коэффициента. В исследовании [36] на нескольких выборках было показано, что определение семантической близости с помощью BERT-подобных нейросетевых моделей сильно коррелирует с экспертными оценками семантической близости. Все это говорит о том, что трансформерные модели на данный момент являются наиболее точным способом моделирования семантической близости текстов.

Сравнение иерархий. На сегодняшний день существует немало способов сравнения иерархий, которые применяются в различных сферах, где возникает задача определения близости деревьев — например, в вычислительной биологии [37, 38]. Из широко используемых метрик можно выделить расстояние Робинсона-Фоулдса [39], коэффициент Жаккара [40], расстояние редактирования [13] и другие. Остановимся в данной работе на последней из рассмотренных метрик, расстоянии редактирования (TED), как одной из наиболее широко применяемых функций расстояния между деревьями [41].

Расстояние редактирования было впервые предложено в работе [13] как минимальная стоимость операций редактирования дерева (добавления, удаления и обновления вершины) для получения одного дерева из другого при заданной стоимости операций редактирования. Авторы, К. Чжан и Д. Шаша, вместе с новой метрикой предложили также алгоритм ее эффективного вычисления для упорядоченных деревьев, который ныне в литературе носит имя авторов. В работе [42] авторы рассмотрели случай неупорядоченных деревьев; ими было показано, что в случае неупорядоченных деревьев задача поиска расстояния редактирования становится NP-сложной, однако для деревьев небольшого размера модификация алгоритма Чжана-Шаша все еще остается применимой. Данная метрика выбрана в качестве основы предложенной в данной работе, TTED, из-за ее интерпретируемости, широкой изученности и универсальности в силу произвольности выбора стоимостей для операций редактирования деревьев.

Оценивание иерархической суммаризации. Стандартным подходом к оценке качества генерации иерархических сводок, как и в целом в суммаризации, является сравнение полученной сводки со сводкой, созданной по тому же документу экспертом. В силу того, однако, что число работ по теме иерархической суммаризации на данный момент невелико, общепринятой метрики для сравнения текстовых иерархий не существует, что затрудняет сравнение различных подходов к решению задачи иерархической суммаризации. Другим распространенным методом оценивания качества суммаризации являются исследования пользовательского опыта, и подобные методы оценивания качества иерархической суммаризации были применены в работах [4] и [5], однако для их применения требуется привлечение большого числа пользователей, из чего следует затратность и низкая воспроизводимость.

В существующих автоматических подходах к оцениванию иерархической суммаризации текстовые деревья, как правило, сравниваются отдельно по своей структуре как деревья и отдельно с помощью, например, метрики ROUGE как наборы текста [6, 7]. Такой подход, однако, не учитывает взаимосвязь между структурой текстового дерева и его содержанием, а применение статистических метрик по типу ROUGE все еще может не учитывать семантические сходства/различия текста [29]. Низкая информативность метрики, использованной в [6], будет показана ниже. Это и отсутствие других воспроизводимых метрик для задачи иерархической суммаризации обуславливает необходимость разработки новой метрики для данной задачи.

3 Предлагаемый метод

3.1 Метрика для сравнения текстовых деревьев

Проанализируем требования к метрике для сравнения текстовых деревьев, сформулированные в разделе 1.3. Из последнего условия (3) естественным образом следует, что расстояние ρ будет соответствовать наименьшему по стоимости набору операций редактирования дерева, так как в противном случае последнее условие можно будет тривиально нарушить. Также стоит отметить, что, во-первых, заданная таким образом функция ρ будет являться метрикой, коль скоро метрикой является $f\left(r(\cdot,\cdot)\right)$, и, во-вторых, заданным требованиям тривиально удовлетворяет расстояние редактирования деревьев (tree edit distance, TED) со стоимостями операций редактирования, заданными в соответствии с выдвинутыми требованиями. Таким образом, определим новую метрику на множестве текстовых деревьев — *расстояние редактирования текстовых деревьев*, или *TTED* (text tree edit distance).

Для вычисления TTED будем применять алгоритм Чжана-Шаша, а именно его модификацию для неупорядоченных деревьев [42]. В качестве стоимости обновления вершины, то есть замены фрагмента текста в вершине на другой, исходя из требований к метрике выше естественно использовать степень семантического сходства этих фрагментов. Для этого предлагается использовать в качестве оценки расстояние между представлениями (эмбеддингами) данных текстов, полученными с помощью заранее выбранной языковой модели.

Пусть имеется языковая модель $\mathrm{LM} : S \to \mathbb{R}^n$, сопоставляющая фрагментам текста некоторые конечномерные эмбеддинги. Тогда мы можем определить для $s, s' \in \mathcal{S}$ семантическое расстояние как $r(s, s') = \rho_n(\mathrm{LM}(s), \mathrm{LM}(s'))$, где ρ_n — функция расстояния в $\mathbb{R}^n$. В качестве меры семантической близости эмбеддингов можно использовать, например, косинусный коэффициент (cosine similarity) S_C, как предлагается в [35]. В таком случае функцию расстояния естественно определить как $\rho_n(A, B) = \sqrt{1 - S_C(A, B)}$. Также в данной работе будет исследовано применение в качестве ρ_n стандартных L_1- и L_2-метрик.

Модификации алгоритма Для улучшения качества TTED и ее эффективного вычисления предлагаются следующие эвристики:

1. **Использование контекста.** Зачастую на практике некорректно сравнивать тексты в вершинах дерева без учета их контекста. Например, предложения «В статье рассказывается про него.» и «В статье рассказывается про метод сравнения текстовых деревьев.» фактически эквивалентны, если в родительской вершине первого стоит предложение «Предлагается новый метод сравнения текстовых деревьев.». В связи с этим в предложенной реализации добавляется возможность при сравнении деревьев предварительно добавлять в вершины все предложения из родительских вершин в качестве контекста перед текстом в вершине и после сравнивать эмбеддинги, полученные с помощью модели с учетом этого контекста.

2. **Предварительное вычисление.** Многократное вычисление эмбеддингов с помощью нейросетевой модели может быть очень затратно по времени для больших деревьев, поэтому предлагается предварительно вычислить эмбеддин-

ги для всех текстов в вершинах и применять предложенный алгоритм уже для дерева из эмбеддингов с вышеуказанной стоимостью обновления меток.

Базовый метод. Для сравнения с предложенной метрикой рассмотрим функцию сходства текстовых деревьев, использованную в работах [6–8] для оценки сходства автоматически сгенерированных иерархических сводок с эталонными. Для текстовых деревьев $T = (V, E)$ и $T' = (V', E')$ функция сходства определяется как:

$$\mathrm{Sim}(T, T') = \min_{P \subset E \times E'} \sum_{(e,e') \in P} \left(\mathrm{ROUGE}(e_0, e'_0) + \mathrm{ROUGE}(e_1, e'_1)\right).$$

где P — однозначное сопоставление ребер T ребрам T', подбираемое жадным алгоритмом, $\mathrm{ROUGE}(v, v')$ — усредненная оценка ROUGE-1, ROUGE-2 и ROUGE-L [12] сходства $s(v)$ и $s(v')$:

$$\mathrm{ROUGE}(s, s') = \frac{1}{3}\left(\mathrm{ROUGE\text{-}1}(s, s') + \mathrm{ROUGE\text{-}2}(s, s') + \mathrm{ROUGE\text{-}L}(s, s')\right).$$

Следует отметить, что базовый метод является функцией сходства, а не метрикой на $\mathcal{T}$. Поскольку TTED является метрикой на $\mathcal{T}$, то для сравнения этих двух методов следует построить некоторую метрику по $\mathrm{Sim}(\cdot, \cdot)$. Это можно сделать, построив псевдометрику по аналогии с ядерным методом [43]:

$$\rho_{\text{baseline}}(T, T') = \sqrt{\mathrm{Sim}(T, T) + \mathrm{Sim}(T', T') - \mathrm{Sim}(T, T') - \mathrm{Sim}(T', T)}.$$

Такое определение базовой метрики, во-первых, автоматически гарантирует следующие свойства:

- Симметричность: $\forall T, T' \in \mathcal{T}\ \rho_{\text{baseline}}(T, T') = \rho_{\text{baseline}}(T', T)$;
- Неотрицательность: $\forall T, T' \in \mathcal{T}\ \rho_{\text{baseline}}(T, T') \geq 0$, причем $\rho_{\text{baseline}}(T, T') = 0 \Leftrightarrow T = T'$.

Во-вторых, неравенство треугольника для ρ_{baseline} также будет выполняться, коль $\mathrm{Sim}(\cdot, \cdot)$ является положительно определенным ядром, однако проверка этого факта в данной работе будет опущена.

3.2 Оценивание качества метрик

Качество метрики ρ на заданном множестве текстовых деревьев $\mathcal{T}$ в постановке задачи (7) будет определяться по коэффициентам $R_S(\rho)$ и $R_M(\rho)$. Очевидно, что вычислить эти величины в точности не представляется возможным, поскольку перебор всех возможных текстовых деревьев является невыполнимой задачей даже в классе деревьев с заданной максимальной глубиной. Для решения этой проблемы можно воспользоваться оценками, полученными с помощью семплирования деревьев T из $\mathcal{T}$ и их модификаций из $P(T)$, $S(T)$ и $M(T)$.

Рассмотрим выборку $\mathcal{D} = \{T, T'_1, \ldots, T'_p, T''_1, \ldots, T''_s, T'''_1, \ldots, T'''_m\}$, где $T \sim \mathcal{T}$, $T'_i \sim P(T)$, $T''_j \sim S(T)$, $T'''_k \sim M(T)$. Введем следующие оценки на $R_S(\rho)$ и $R_M(\rho)$ по $\mathcal{D}$:

$$R_S^{\mathcal{D}}(\rho) = \frac{1}{sp}\sum_{i=1}^{p}\sum_{j=1}^{s}\frac{\rho(T, T'_i)}{\rho(T, T''_j)}, \quad R_S^{\mathcal{D}}(\rho) = \frac{1}{mp}\sum_{i=1}^{p}\sum_{k=1}^{m}\frac{\rho(T, T'_i)}{\rho(T, T'''_k)}.$$

Положим, что введенные в постановке задачи величины $r_S(\rho, T)$, $r_M(\rho, T)$, $R_S(\rho)$, $R_M(\rho)$ корректны для выбранной метрики ρ и класса текстовых деревьев $\mathcal{T}$. Тогда нетрудно видеть, что оценки $R_S(\rho)$ и $R_M(\rho)$ по выборке $\mathcal{D}$ будут несмещенными:

Теорема 1 (Соболевский, 2025) *Пусть для заданного класса текстовых деревьев $\mathcal{T}$ и метрики $\rho : \mathcal{T} \times \mathcal{T} \longrightarrow [0, +\infty)$ существуют конечные $R_S(\rho)$ и $R_M(\rho)$. Тогда $R_S^{\mathcal{D}}(\rho)$ и $R_M^{\mathcal{D}}(\rho)$ являются несмещенными оценками $R_S(\rho)$ и $R_M(\rho)$ соответственно по выборке $\mathcal{D}$:*

$$\mathbb{E}_{\mathcal{D}}[R_S^{\mathcal{D}}(\rho)] = R_S(\rho), \quad \mathbb{E}_{\mathcal{D}}[R_M^{\mathcal{D}}(\rho)] = R_M(\rho).$$

Доказательство. Здесь приводится доказательство для $R_S^{\mathcal{D}}(\rho)$, однако доказательство для второй оценки является точно таким же с точностью до переобозначения. Распишем математическое ожидание по выборке $\mathcal{D}$ через математические ожидания по ее элементам, пользуясь свойством условного математического ожидания:

$$\mathbb{E}_{\mathcal{D}}[R_S^{\mathcal{D}}(\rho)] = \mathbb{E}_{T\sim\mathcal{T}}\left[\mathbb{E}_{T_i'\sim P(T),\ T_j''\sim S(T)}\left[\frac{1}{sp}\sum_{i=1}^{p}\sum_{j=1}^{s}\frac{\rho(T,T_i')}{\rho(T,T_j'')}\middle| T\right]\right].$$

Далее, из линейности математического ожидания

$$\mathbb{E}_{T_i'\sim P(T),\ T_j''\sim S(T)}\left[\frac{1}{sp}\sum_{i=1}^{p}\sum_{j=1}^{s}\frac{\rho(T,T_i')}{\rho(T,T_j'')}\middle| T\right] = \frac{1}{sp}\sum_{i=1}^{p}\sum_{j=1}^{s}\mathbb{E}_{T_i'\sim P(T),\ T_j''\sim S(T)}\left[\frac{\rho(T,T_i')}{\rho(T,T_j'')}\middle| T\right].$$

Несложно видеть, что под знаком суммы стоит ничто иное, как выражение для $r_S(\rho, T)$. Следовательно,

$$\mathbb{E}_{\mathcal{D}}[R_S^{\mathcal{D}}(\rho)] = \frac{1}{sp}\sum_{i=1}^{p}\sum_{j=1}^{s}\mathbb{E}_{T\sim\mathcal{T}}[r_S(\rho,T)] = \mathbb{E}_{T\sim\mathcal{T}}[r_S(\rho,T)] = R_S(\rho),$$

где последнее равенство есть просто определение $R_S(\rho)$. ∎

Следует отметить, что несмещенность данной оценки не зависит от вида распределений вероятностей семплирования из множеств $\mathcal{T}$ и $P(T)$, $S(T)$, $M(T)$, что позволяет избежать дополнительного исследования этих распределений, чего и хотелось добиться от оценки на коэффициенты $R_s(\rho)$ и $R_M(\rho)$. Имея выборку $\mathcal{D}$ обозначенного выше вида и несмещенные оценки данных коэффициентов по ней, можно переписать оптимизационные задачи (7) в следующем виде:

$$R_S^{\mathcal{D}}(\rho) \longrightarrow \min_{\rho}, \quad R_M^{\mathcal{D}}(\rho) \longrightarrow \min_{\rho}. \tag{8}$$

3.3 Многокритериальное сравнение текстовых деревьев

Использование агрегированной метрики сходства удобно тем, что упрощает задачу оптимизации генерации текстовых деревьев, однако оно не является способом получить информацию об отличии и сходстве генерируемых иерархических сводок с авторскими по каждому аспекту сходства отдельно. В связи с этим воспользуемся несколькими метриками сходства, каждая из которых будет отражать аспекты сходства своей природы. Основными аспектами, по которым будут сравниваться текстовые деревья, будут:

1. *Семантика* дерева, то есть его смысловое содержание в отрыве от древовидной структуры;

2. *Структура* дерева без учета текстовых меток вершин.

3. *Ранжирование* текстов в иерархии.

Последний из выделенных аспектов сходства текстовых деревьев в рамках данной работы рассматриваться не будет.

Семантическое сходство. Первый измеряемый в данной работе аспект сходства предполагает сравнение текстов в вершинах деревьев по смыслу как наборов предложений без графовой структуры. Здесь целесообразно применить одну из метрик, используемых для оценки качества обычной, линейной суммаризации. Среди метрик сходства текста наиболее информативными по определению семантической близости являются методы, основанные на нейросетевом моделировании текстов в виде векторов. Применим одну из таких метрик — *BERTScore* [30], основанную на сравнении текстов по их представлениям, полученным с помощью BERT-подобной языковой модели.

Структурное сходство. Второй аспект, по которому можно сравнивать иерархические сводки — их структура. Для сравнения структур текстовых деревьев можно сравнивать их как неразмеченные деревья, игнорируя текст в вершинах. Существуют разные метрики сравнения иерархий, однако остановимся на метрике, которая уже была применена в данной работе и нашла широкое применение для сравнения деревьев — *расстоянии редактирования* [13]. В данном случае способом сравнить исключительно структуры деревьев будет применить алгоритм поиска расстояния редактирования для них, задав лишь стоимости операций удаления и добавления вершины равными фиксированному числу — например, 1. Тогда заданное таким образом расстояние редактирования будет показывать, сколько операций редактирования минимально необходимо совершить, чтобы получить из структуры первого дерева структуру второго.

3.4 Генерация иерархических сводок с помощью БЯМ

Основной метод работы с БЯМ — создание запросов к модели под конкретную задачу, т. е. *промптинг*. Выделим два основных метода генерации иерархической сводки с помощью промптинга БЯМ:

1. Прямой промптинг — генерация иерархической сводки моделью полностью по одному запросу, содержащему текст документа;

2. Последовательный промптинг — последовательная генерация вершин сводки несколькими запросами с участием пользователя в определении траектории генерации.

Прямой промптинг. Данный метод подразумевает генерацию иерархической сводки сразу полностью с использованием запроса, состоящего из некоторого общего для задачи шаблона запроса и текста документа, который требуется суммаризировать. Этот метод является прямолинейным и распространенным методом использования БЯМ для различных задач генерации текста — в частности, для обычной суммаризации текстов [3]. Несмотря на его простоту и небольшое время работы, он имеет один недостаток — сильную зависимость от содержания шаблона запроса. Результаты, получаемые с помощью БЯМ, сильно варьируются в зависимости от вида запроса к ней, и для каждой модели приходится подбирать свои оптимальные запросы. Потенциальным решением данной проблемы является автоматическая оптимизация запросов, однако, как можно будет увидеть дальше, в задаче иерархической суммаризации результаты, близкие к авторским, можно получить поиском по набору созданных вручную запросов к модели.

Последовательный промптинг. Второй метод подразумевает взаимодействие с БЯМ посредством последовательных запросов на генерацию новых вершин дерева иерархической сводки. Данный метод представляется способом суммаризировать информацию из научной статьи в виде структуры, созданной под запросы пользователя. Предлагается следующий алгоритм взаимодействия с системой:

1. Пользователь загружает документ, который требуется изучить, в систему;
2. БЯМ генерирует корневую вершину сводки в соответствии с запросом (например, выделяет основную мысль текста) и ряд уточняющих вопросов, предлагающих пользователю узнать подробнее про разные аспекты документа;
3. Пользователь выбирает вопросы, на которые он хочет получить ответы по тексту документа, либо завершает процесс генерации сводки и переходит к шагу 7;
4. БЯМ генерирует ответы на вопросы, выбранные пользователем, и добавляет их в иерархическую сводку в качестве дочерних вершин той вершины, к которой были сгенерированы уточняющие вопросы;
5. Пользователь выбирает вершину сводки, по тексту в которой хочет узнать больше деталей, либо завершает процесс генерации сводки и переходит к шагу 7;
6. БЯМ генерирует уточняющие вопросы к выбранной вершине, и алгоритм переходит на шаг 3;
7. Когда генерация сводки завершена, система собирает сводку и экспортирует ее в нужном пользователю формате.

Соответственно, для использования БЯМ для иерархической суммаризации по такому алгоритму необходимы следующие шаблоны запросов:

- Запрос на изначальную обработку текста документа и генерацию корневой вершины дерева сводки с указанием того, что должно являться корнем дерева;
- Запрос на генерацию уточняющих вопросов к вершине дерева;

- Запрос на генерацию новых вершин как ответов на выбранные пользователем вопросы;
- Возможно, запрос на корректировку формата генерации.

Такое количество разных запросов для оптимизации является недостатком данного метода вместе с его вычислительной сложностью. Также стоит учесть, что данный метод в такой его реализации не является автоматическим и предполагает участие пользователя. Несмотря на это, такой способ имеет и очевидные преимущества: автоматический учет предпочтений пользователя по содержанию и структуре генерируемой иерархической сводки и интерактивность процесса иерархической суммаризации. Качество иерархической суммаризации таким методом логичнее всего измерять, как максимальное по всем сценариям взаимодействия пользователя с системой сходство полученной иерархической сводки с авторской.

4 Вычислительные эксперименты

4.1 Тестирование метрики для сравнения текстовых деревьев

Постановка эксперимента. Для проверки применимости предложенной метрики сходства текстовых деревьев в сравнении с базовым методом проведем вычисление оценок расстояния на выборке $\mathcal{D}$, состоящей из следующих элементов:

1. текстового дерева T;

2. деревьев, которые идентичны по семантике и структуре, но тексты в узлах дерева *перефразированы* — подвыборка $\mathcal{T}_1$ из $P(T)$;

3. деревьев, которые сформированы из одних и тех же вершин, но с разной *структурой* дерева — подвыборка $\mathcal{T}_2$ из $S(T)$;

4. деревьев, которые идентичны по структуре и схожи по наборам слов в текстах вершин, но значительно *отличаются по значению* — подвыборка $\mathcal{T}_3$ из $M(T)$.

Цель — найти среди предложенных такую метрику ρ, для которой будут минимальными оценки $R_S^{\mathcal{D}}(\rho)$ и $R_M^{\mathcal{D}}(\rho)$ согласно задачам минимизации (8). Введем также обозначение $\overline{\rho}_i$ для среднего расстояния между T и деревьями из подвыборки $\mathcal{T}_i$. Тогда качественным признаком информативности метрики ρ будет значительное отличие в меньшую сторону значения $\overline{\rho}_1$ от значений $\overline{\rho}_2$ и $\overline{\rho}_3$.

Экспериментальные данные. Поскольку привлечь других экспертов или краудсорсинг к созданию выборки для данной работы не представилось возможным, а ручная иерархическая суммаризация занимает довольно много времени и не обеспечивает при этом разнообразности данных, для создания выборки для тестирования метрик на чувствительность к различным аспектам сходства текстовых деревьев была привлечена генеративная нейросеть DeepSeek V3 [44]. Процесс генерации данных состоял из следующих этапов:

1. Создание базовой иерархической сводки по научному исследованию с помощью нейросети;

2. Генерация модификаций этой сводки с помощью нейросети;

3. Ручная проверка сгенерированных текстовых деревьев на соответствие требованиям к выборке.

Такая методика генерации данных позволила значительно сократить затрачиваемое на создание выборки время в условиях самостоятельной работы и избежать смещения в данных, обусловленного работой одного эксперта, при этом сохраняя контроль качества искусственно сгенерированных данных. В результате была получена выборка, состоящая из базового дерева и десяти его модификаций в каждой категории модификаций. Запросы, при помощи которых создавались текстовые деревья для выборки, как и сама выборка, приведены в репозитории проекта: `https://github.com/intsystems/Sobolevsky-BS-Thesis`.

Реализация. Алгоритм Чжана-Шаша для вычисления TTED был реализован при помощи Python-библиотеки `zss`. Языковые модели для моделирования семантической близости текстов были взяты из библиотеки `sentence-transformers`. Были применены следующие языковые модели:

- Дообученная DistilRoBERTa (`paraphrase-distilroberta-base-v1`)[5];
- SPECTER (`allenai-specter`) [45];
- MPNet (`all-mpnet-base-v2`) [46];
- Дообученная MPNet (`paraphrase-multilingual-mpnet-base-v2`)[6].

Реализация базового метода взята из официального репозитория статьи [6]: `https://github.com/Cyno2232/CMGN`. Данная реализация была адаптирована под формат генерации в данном эксперименте, однако основная логика использованного метода была оставлена в изначальном виде. Весь код, позволяющий воспроизвести полученные в данном эксперименте результаты, доступен в репозитории исследования.

Результаты. Результаты тестирования различных языковых моделей с расстоянием на основе косинусного коэффициента в сравнении с базовым методом представлены в таблице 1. Оценки расстояния, полученные с помощью базового метода и TTED с использованием дообученной модели MPNet представлены на рис. 1a и 1b соответственно.

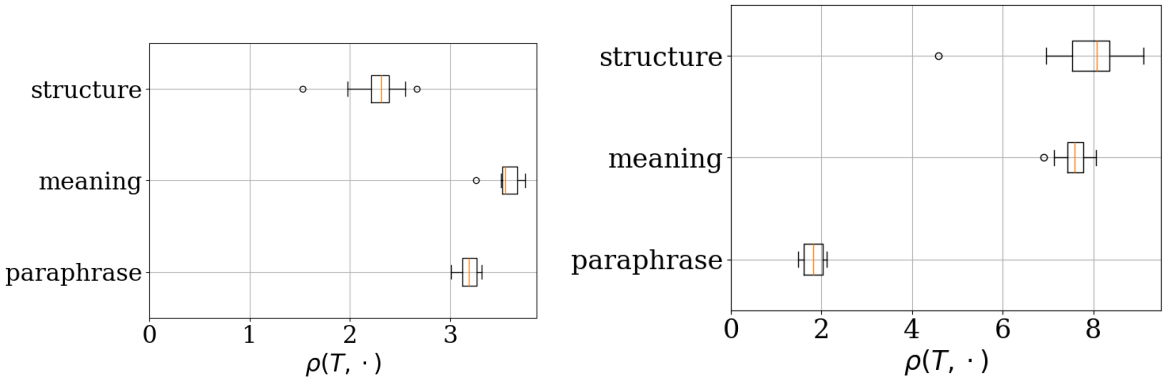
Таблица 1: Средние оценки расстояния с помощью разных языковых моделей

Модель	$\overline{\rho}_1$	$\overline{\rho}_2$	$\overline{\rho}_3$	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Базовый метод	3,18±0,09	2,26±0,30	3,56±0,13	1,44±0,25	0,89±0,03
DistilRoBERTa	3,33±0,23	7,76±1,74	7,38±0,41	0,46±0,14	0,45±0,04
SPECTER	1,39±0,16	3,70±0,81	4,74±0,66	0,40±0,12	0,30±0,06
MPNet	2,30±0,33	7,19±1,25	8,06±0,67	0,33±0,08	0,29±0,03
Дообученная MPNet	1,82±0,22	7,71±1,18	7,56±0,33	**0,24±0,05**	**0,24±0,03**

По рис. 1b и значениям $R_2(\rho)$ в таблице 1 видно, что TTED показывает значения расстояния значительно больше для деревьев, различающихся по семантике и структуре, чем для деревьев, которые являются перефразированием друг друга. Это подкрепляет утверждение о том, что TTED отражает значимые отличия текстовых деревьев заметно сильнее, чем незначительные. Более того, для большинства исследованных энкодеров, включая показавшийся лучшие результаты, расстояния между отличающимися по структуре и семантике деревьями в среднем оказались почти равны. Это позволяет утверждать, данная метрика сбалансирована по различным аспектам качества.

Для сравнения рассмотрим значения, полученные с помощью базового метода (рис. 1a, таблица 1). Можно видеть несколько иную картину: прежде всего, структурные различия с помощью данного метода отражаются гораздо слабее, чем

[5] `https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v1`
[6] `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`



(a) Оценки базового метода (b) Оценки TTED с дообученной MPNet

Рис. 1: Оценки расстояний по разным аспектам различия

отличия текстов в вершинах, в том числе по формулировкам, поэтому значение $R_S(\rho_{\text{baseline}}) > 1$. В случае же значимых семантических различий отличие расстояний от расстояний между парафразами в среднем, хоть и в большую сторону, относительно невелико, что отражено в заметно большем значении коэффициента R_M. Здесь, конечно, стоит сделать оговорку: базовый метод изначально — функция сходства, а функция расстояния по нему построена автором специально для данного эксперимента. Однако же посмотрим на средние значения $\overline{\text{Sim}_i}$ исходной функции сходства $\text{Sim}(\cdot, \cdot)$ на рассмотренных выборках $\mathcal{T}_i$ соответственно:

$$\overline{\text{Sim}_1} = 3,92 \pm 0,29, \quad \overline{\text{Sim}_2} = 6,92 \pm 0,67, \quad \overline{\text{Sim}_3} = 2,64 \pm 0,46.$$

По этим значениям также видно, что разница между оценками сходства для первой и третьей выборками также невелика (меньше, чем в два раза). Для сравнения, средние значения расстояний, полученные с помощью TTED на тех же выборках, отличаются больше чем в четыре раза, что является гораздо более значительным отличием. Все вышеперечисленное подтверждает гипотезу о том, что предложенный метод TTED заметно информативнее базового метода.

На примере версии TTED с дообученной языковой моделью MPNet были проведены эксперименты с разными эмбеддинговыми расстояниями и без использования контекста при вычислении эмбеддингов. Результаты этих экспериментов представлены в таблицах 2 и 3. Видно, что тип используемой для измерения расстояния между эмбеддингами метрики не сильно влияет на итоговый коэффициент качества отражения семантической близости, однако значения расстояния сильно различаются по порядку величины в зависимости от метрики. В дальнейших экспериментах предпочтительной будет метрика, основанная на косинусной близости, так как она дает наиболее интерпретируемые значения (с ее использованием расстояние между эмбеддингами является числом от 0 до 1). По таблице 3 видно, что использование контекста как эвристика для алгоритма вычисления TTED действительно дает небольшое улучшение информативности метрики, что обосновывает ее использование в дальнейших экспериментах.

$r(x,y)$	$\overline{\rho}_1$	$\overline{\rho}_3$	$R_M^{\mathcal{D}}(\rho)$
$\sqrt{1-S_C(x,y)}$	1,82±0,22	7,56±0,33	**0,24±0,03**
$\|\|x-y\|\|_2$	7,34±0,92	30,22±1,31	**0,24±0,03**
$\|\|x-y\|\|_1$	157,09±19,87	617,63±27,16	0,25±0,03

Таблица 2: Средние значения TTED для разных эмбеддинговых расстояний

Метод	$\overline{\rho}_1$	$\overline{\rho}_2$	$\overline{\rho}_3$	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Без контекста	1,25±0,10	4,80±1,63	4,34±0,52	0,32±0,19	0,29±0,02
С контекстом	1,82±0,22	7,71±1,18	7,56±0,33	**0,24±0,05**	**0,24±0,03**

Таблица 3: Зависимость расстояний от использования контекста

4.2 Иерархическая суммаризация с помощью БЯМ

Постановка эксперимента. Для качественного анализа способности БЯМ к иерархической суммаризации текстов научных публикаций в рамках данного исследования проводится эксперимент с одной из современных БЯМ, моделью Mistral Large 2, в применении к иерархической суммаризации выборки из пяти научных текстов с помощью предложенных в разделе 3.4 методов промптинга. Для каждого из данных методов тестируется ряд сценариев работы с БЯМ, позволяющих получить корректную иерархическую сводку научного документа, и определяется оптимальный по расстоянию по метрике TTED от авторской сводки сценарий. В качестве дополнительных метрик для оценки качества иерархической суммаризации используются метрики семантического сходства и структурного различия, предложенные в разделе 3.3. Цель данного эксперимента — оценить качество генерации иерархических сводок с помощью БЯМ с применением различных стратегий работы с моделью, а также применить на практике предложенную в данном исследовании метрику для сравнения текстовых деревьев.

Экспериментальные данные. Для тестирования иерархической суммаризации научных статей с помощью БЯМ в данной работе используется пять разных научных статей на английском языке из списка используемой литературы. По каждой статье была вручную построена иерархическая сводка в качестве авторской для сравнения с генерируемыми. В рамках данного эксперимента для единообразия в качестве авторских строились иерархические сводки из предложений, содержащие ровно три уровня иерархии и главную мысль текста в качестве корневой вершины; в запросах к БЯМ давались соответствующие указания по глубине генерируемых деревьев.

Реализация. Работа с БЯМ Mistral Large 2 была реализована через официальный пользовательский интерфейс La Plateforme[7]. Реализация метрики BERTScore была взята из Python-библиотеки `bert-score` [30]. В метрике TTED в качестве языковой модели использовалась дообученная модель MPNet с метрикой на основе косинусной близости как наиболее информативная по результатам предыдущего эксперимента.

Для прямого промптинга было вручную создано четыре различных шаблона запросов к БЯМ, и по результатам тестирования каждого из них определялся оп-

[7] `https://mistral.ai/news/mistral-large-2407`

Метод промптинга	BERTScore	TED	TTED
Прямой	**0,87±0,01**	3,80±3,76	7,46±3,11
Последовательный	**0,88±0,01**	**0,00**	**3,76±0,39**

Таблица 4: Результаты тестирования БЯМ для иерархической суммаризации

тимальный. Для метода последовательного промптинга также вручную были подобраны текстовые запросы, позволяющие корректно реализовать алгоритм иерархической суммаризации с помощью данного метода. Все использованные в данном эксперименте шаблоны запросов и код, позволяющий воспроизвести данный эксперимент, доступны в репозитории исследования. Следует отметить только, что в точности полученные результаты воспроизвести скорее всего не удастся в силу фактора случайности в реализации выбранной БЯМ.

Результаты. Средние значения метрик сходства/различия сгенерированных с помощью БЯМ иерархических сводок с авторскими по выборке представлены в таблице 4. Здесь приведены лучшие результаты, которые позволил получить каждый из методов: для прямого промптинга это результаты для оптимального (по метрике TTED) шаблона запроса, для последовательного — средние результаты для оптимальных сценариев взаимодействия с системой.

По результатам в таблице 4 видно, что семантическое сходство сгенерированных сводок с авторским в среднем примерно одинаково высоко. Полученные значения BERTScore сравнимы со значениями для самых современных моделей суммаризации текстов [47, 48], что еще раз подтверждает высокую способность БЯМ к суммаризации текстов, причем из этого небольшого эксперимента видно, что запрос на структурирование информации не сказывается на качестве передачи семантики исходного текста.

Ключевое различие двух исследованных методов заключается в сходстве авторских и потенциально генерируемых с помощью каждого метода сводок по структуре: метод последовательного промптинга позволяет получить сводки с минимальным различием по структуре между сгенерированной и авторской сводкой. Это неудивительно, поскольку при таком алгоритме взаимодействия с системой оптимальным по метрике TTED сценарием во всех случаях оказывается именно тот, по которому строится иерархическая сводка такой же структуры, что и авторская. За счет этого с помощью данного метода получилось добиться в среднем вдвое меньшего общего расстояния между сгенерированными и авторскими сводками, чем с помощью прямого промптинга. Для последнего метода структурное сходство сгенерированных сводок с авторскими сильно варьируется и может достигать достаточно высоких значений при неправильной трактовке моделью требований к структуре. Одним из наблюдений, полученных при проведении экспериментов с прямым промптингом, было то, что модели для получения оптимального по структуре результата необходимо как можно более четкое указание того, какая структура сводки от нее требуется. Даже при оптимальном промптинге, однако, идеальное совпадение структур на всей выборке получить не удалось, поскольку иерархические сводки в ней, хоть и довольно схожие по структуре, все же имеют небольшие различия, которые модель не может воспроизвести идеально.

Заключение

Обсуждение результатов. В данной работе была предложена новая метрика качества иерархической суммаризации — TTED. С помощью исследования данной метрики в сравнении с существующей по чувствительности к значимым различиям иерархических сводок было показано, что предложенная метрика действительно превосходит по информативности применяемую в предыдущих работах по теме. При грамотном выборе модели-кодировщика для аппроксимации семантической близости в TTED в данной работе удалось получить значительно лучшие значения предложенных коэффициентов качества, чем с помощью базового метода; качественное исследование полученных результатов также указывает на превосходство TTED перед базовым методом. Все это указывает на применимость метрики TTED в дальнейших работах по теме иерархической суммаризации, а также потенциально для сравнения текстовых деревьев в других целях.

Применение БЯМ в задаче иерархической суммаризации также показало многообещающие результаты: результаты в терминах семантического сходства полученных иерархических сводок с авторскими сопоставимы с результатами для лучших из современных систем машинной суммаризации текстов, а при правильно подобранной стратегии промптинга модели можно также минимизировать различие сводок по структуре и гарантировать генерацию моделью структурированных сводок в корректном формате для дальнейшего сравнения и обработки. С этой целью оптимальнее использовать метод последовательного промптинга как способ лучше подстроиться под запросы пользователя к иерархической сводке, однако метод прямого промптинга при удачном выборе шаблона запроса также может показывать результаты, близкие к человеческим. В конечном итоге выбор стратегии взаимодействия с БЯМ зависит от цели генерации, будь то автоматическая быстрая генерация потенциально нескольких сводок сразу либо интерактивное получение информации под специфические запросы пользователя.

Вышеуказанные результаты исследования позволяют утверждать, что, во-первых, для задачи иерархической суммаризации существует, по меньшей мере, сильный базовый метод — иерархическая суммаризация с помощью БЯМ, с которым можно сравнивать другие методы, и, во-вторых, что теперь существует способ информативно оценивать генерацию иерархических сводок. Данная работа может стать основой для дальнейших исследований по теме иерархической суммаризации и призвана подогреть интерес научного сообщества к этой теме.

Направления дальнейшей работы. У представленного исследования есть ряд ограничений, который в дальнейшем планируется устранить. Во-первых, данная работа является индивидуальной, что накладывает ограничения как по человеческим ресурсам, требуемым для создания и верификации данных для тестирования предложенных методов, так и по объективности авторской иерархической суммаризации. Одной из важных перспектив исследования является создание выборки для задачи большего размера силами нескольких экспертов и сравнение степени различия между экспертными и сгенерированными сводками со средней степенью различия между экспертными сводками, а также создание «золотых стандартов» в виде иерархических сводок, созданных совместным трудом сразу нескольких экспертов. Такой эксперимент позволит точнее определить реальное качество генерации иерархических

сводок с помощью БЯМ.

Эксперименты, проведенные с БЯМ в рамках данного исследования, имеют качественный характер и не покрывают всего многообразия способов и сценариев использования БЯМ для иерархической суммаризации. В частности, в данной работе в качестве научных текстов рассмотрен только ряд статей по темам, близким к теме данного исследования, однако область применения иерархической суммаризации может не ограничиваться суммаризацией даже только лишь научных текстов — эта область выбрана в рамках данного исследования для иллюстрации одного из наиболее практически полезных, по мнению автора, сценариев применения иерархической суммаризации. В рамках экспериментов исследовалась только одна модель, Mistral Large 2, и в рамках будущей работы планируются в том числе исследования большего числа моделей.

Есть также некоторые перспективы исследований, связанные с предложенной метрикой и многокритериальным сравнением текстовых деревьев. Во-первых, существует больше аспектов сходства текстовых деревьев, чем только лишь структура и семантика. Данные аспекты планируется формализовать и исследовать в дальнейшем. Также следует учесть, что асимптотически алгоритм Чжана-Шаша, особенно для неупорядоченных деревьев, является достаточно ресурсозатратным и потенциально может сделать TTED неоптимальной метрикой для больших текстовых деревьев; границу ее применимости в терминах размера сравниваемых деревьев еще предстоит определить. Данное исследование в экспериментах ограничивается деревьями глубины 3 примерно одинаковой структуры, но, очевидно, в перспективе можно изучить генерацию и сравнение деревьев с различными размерами и структурой.

Данное исследование сфокусировано на качестве иерархической суммаризации как на сходстве машинной суммаризации с человеческой, однако отдельной большой перспективой будущих исследований является оценивание иерархических сводок с использованием только лишь текста документа либо самих по себе в терминах, например, полноты представления текста, непротиворечивости, связности и избыточности. Также важным направлением дальнейшей работы является исследование качества иерархической суммаризации в пользовательских исследованиях для определения практической применимости данного метода.

Основные положения, выносимые на защиту.

- Введен новый коэффициент качества метрик на множестве текстовых деревьев, позволяющий оценить информативность метрики как функции расстояния, учитывающей их структуру и семантику, и предложена несмещенная оценка данного коэффициента по выборке текстовых деревьев.

- Разработан новый алгоритм оценки расстояния между текстовыми деревьями, позволяющий агрегировать различные аспекты различия текстовых деревьев и лучше отражающий значимые отличия текстовых деревьев в терминах введенного коэффициента качества, чем используемый до этого метод.

- Проведено многокритериальное исследование двух методов иерархической суммаризации при помощи БЯМ с использованием предложенного нового метода сравнения текстовых деревьев как метода сравнения сгенерированных иерархических сводок с экспертными.

Список литературы

[1] Buzan Tony. Use your head. — Pearson Education, 2006.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Beyond generic summarization: A multi-faceted hierarchical summarization corpus of large heterogeneous data / Tauchmann Christopher, Arnold Thomas, Hanselowski Andreas, Meyer Christian M, and Mieskes Margot // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). — 2018.

[10] Litvak Marina, Vanetik Natalia, Puchinsky Zvi. Hierarchical summarization of financial reports with RUNNER // Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation. — 2020. — P. 213–225.

[11] 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.

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

[13] Zhang Kaizhong, Shasha Dennis. Simple fast algorithms for the editing distance between trees and related problems // SIAM journal on computing. — 1989. — Vol. 18, no. 6. — P. 1245–1262.

[14] Summarizing text documents: Sentence selection and evaluation metrics / Goldstein Jade, Kantrowitz Mark, Mittal Vibhu, and Carbonell Jaime // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — 1999. — P. 121–128.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Tungprapa Taviga. Effect of using the electronic mind map in the educational research methodology course for master-degree students in the faculty of education // International Journal of Information and Education Technology. — 2015. — Vol. 5, no. 11. — P. 803.

[22] Rezapour-Nasrabad Rafat. Mind map learning technique: An educational interactive approach // International Journal of Pharmaceutical Research. — 2019. — Vol. 11, no. 1. — P. 1593–1597.

[23] 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.

[24] Direct automatic generation of mind maps from text with M 2 Gen / Abdeen Mohammad, El-Sahan R, Ismaeil A, El-Harouny S, Shalaby M, and Yagoub MCE // 2009 IEEE Toronto international conference science and technology for humanity (TIC-STH) / IEEE. — 2009. — P. 95–99.

[25] Elhoseiny Mohamed, Elgammal Ahmed. English2mindmap: An automated system for mindmap generation from english text // 2012 IEEE International Symposium on Multimedia / IEEE. — 2012. — P. 326–331.

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

[27] 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.

[28] 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.

[29] 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.

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

[31] Bert: Pre-training of deep bidirectional transformers for language understanding / Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina // Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). — 2019. — P. 4171–4186.

[32] Human-like summarization evaluation with chatgpt / Gao Mingqi, Ruan Jie, Sun Renliang, Yin Xunjian, Yang Shiping, and Wan Xiaojun // arXiv preprint arXiv:2304.02554. — 2023.

[33] 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.

[34] 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.

[35] Vrbanec Tedo, Meštrović Ana. Comparison study of unsupervised paraphrase detection: Deep learning—The key for semantic similarity detection // Expert systems. — 2023. — Vol. 40, no. 9. — P. e13386.

[36] Chandrasekaran Dhivya, Mago Vijay. Evolution of semantic similarity—a survey // Acm Computing Surveys (Csur). — 2021. — Vol. 54, no. 2. — P. 1–37.

[37] Lin Yu, Rajan Vaibhav, Moret Bernard ME. A metric for phylogenetic trees based on matching // IEEE/ACM Transactions on Computational Biology and Bioinformatics. — 2011. — Vol. 9, no. 4. — P. 1014–1022.

[38] Pazos Florencio, Valencia Alfonso. Similarity of phylogenetic trees as indicator of protein–protein interaction // Protein engineering. — 2001. — Vol. 14, no. 9. — P. 609–614.

[39] Robinson David F, Foulds Leslie R. Comparison of phylogenetic trees // Mathematical biosciences. — 1981. — Vol. 53, no. 1-2. — P. 131–147.

[40] Jaccard Paul. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines // Bull Soc Vaudoise Sci Nat. — 1901. — Vol. 37. — P. 241–272.

[41] Akutsu Tatsuya. Tree edit distance problems: Algorithms and applications to bioinformatics // IEICE transactions on information and systems. — 2010. — Vol. 93, no. 2. — P. 208–218.

[42] Zhang Kaizhong, Statman Richard, Shasha Dennis. On the Editing Distance Between Unordered Labeled Trees. // Information Processing Letters. — 1992. — 05. — Vol. 42. — P. 133–139.

[43] Schölkopf Bernhard. The kernel trick for distances // Advances in neural information processing systems. — 2000. — Vol. 13.

[44] Deepseek-v3 technical report / Liu Aixin, Feng Bei, Xue Bing, Wang Bingxuan, Wu Bochao, Lu Chengda, Zhao Chenggang, Deng Chengqi, Zhang Chenyu, Ruan Chong, et al. // arXiv preprint arXiv:2412.19437. — 2024.

[45] Specter: Document-level representation learning using citation-informed transformers / Cohan Arman, Feldman Sergey, Beltagy Iz, Downey Doug, and Weld Daniel S // arXiv preprint arXiv:2004.07180. — 2020.

[46] Mpnet: Masked and permuted pre-training for language understanding / Song Kaitao, Tan Xu, Qin Tao, Lu Jianfeng, and Liu Tie-Yan // Advances in neural information processing systems. — 2020. — Vol. 33. — P. 16857–16867.

[47] Kadhim Estabraq Abdulredaa, Feizi-Derakhshi Mohammad-Reza, Aghdasi Hadi S. Advanced Text Summarization Model Incorporating NLP Techniques and Feature-Based Scoring // IEEE Access. — 2025.

[48] Kim Hui-Sang, Kang Ji-Won, Choi Sun-Yong. ChatGPT vs. Human Journalists: Analyzing News Summaries Through BERTScore and Moderation Standards // Electronics. — 2025. — Vol. 14, no. 11. — P. 2115.