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
2025

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 из предложений, или иерархические сводки. Поскольку человеческая обработка больших коллекций текстовых документов, особенно научных, занимает много времени и ресурсов, для решения задачи их иерархической суммаризации необходимо разрабатывать автоматические методы, 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. Поскольку на данный момент не существует выборок для данной задачи, предварительно создается выборка интеллект-карт по ряду научных статей для сравнения со сгенерированными автоматически. Иерархическая суммаризация с помощью БЯМ оценивается в сравнении с сводками из этой выборки с учетом различных аспектов сходства иерархических сводок, таких как структура, семантика, ранжирование фактов в иерархии и другие.

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. С связи с этим в данной работе предлагается также новый метод сравнения текстовых деревьев — расстояние редактирования текстовых деревьев (TTED), основанный 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 относительную чувствительность фукнции сходства к семантическим и структурным различиям текстовых деревьев по отношению к перефразированию предложений в вершинах, а также предлагаются несмещенные оценки на эти коэффициенты по выборкам текстовых деревьев. С помощью этих оценок даётся количественная оценка качества предложенной метрики и ее модификаций в сравнении с использованной до этого.

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

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 одновременным накоплением старых все острее необходимость как в эффективном поиске публикаций, актуальных интересующей теме, так и в эффективном извлечении знаний из них. Поскольку современному исследователю в среднем требуется обработать довольно большой объем научных статей, чтобы получить нужную ему информацию по интересующей теме, возникает потребность в организации информации из публикаций, позволяющей читателю ознакомиться с их содержанием с нужной ему 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 иерархического 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 карт не является новой идеей. В литературе давно известен термин «интеллект-карта» (mind map), обозначающий древовидную карту, иерархически раскрывающую тему от главных понятий к деталям (TO-DO: ссылка на источник). Интеллект-карты показали себя как инструмент, позволяющий улучшить восприятие и запоминание информации [1], однако самостоятельное построение интеллект-карты по тексту научной публикации является весьма времязатратным 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 человекочитаемого 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 [2]. Несмотря на это, хотя в последние годы появилось небольшое число работ, исследующих способность БЯМ генерировать структурированные представления информации [3], 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 [3–7], 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 экс-

трактивным 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.
- Предложить новую агрегированную метрику качества для задачи иерархической суммаризации, отражающую прежде всего значимые различия текстовых иерархий.
- Формализовать требования к адекватности метрики на множестве текстовых иерархий как метрики на множестве объектов, обладающих различными аспектами сходства.
- Разработать и применить методику построения интеллект-карт по научным текстам с целью формирования выборки для обучения и валидации моделей автоматической иерархической суммаризации.
- Применить БЯМ для генерации интеллект-карт по текстам научных статей и определить оптимальный метод работы с моделью, позволяющий максимизировать качество генерации для 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.

Научная новизна. *TO-DO*

to my best knowledge, на данный момент нет ни адекватных критериев для автоматического оценивания карт знаний по научным статьям, ни выборки для данной задачи

Теоретическая значимость. *TO-DO*

- формализация задачи гибридной иерархической суммаризации в виде интеллект-карт из предложений

- разработка автоматических метрик оценки качества иерархических представлений текста по собственной структуре карты и в сравнении с референсом

Практическая значимость. *TO-DO*

- разработка метода автоматической генерации представлений научных статей для их эффективного изучения с нужной степенью углубления в детали

- датасет для задачи генерации интеллект-карт из предложений по научным текстам

- (было бы неплохо?) фреймворк для автоматического оценивания подобных интеллект-карт

Степень достоверности и апробация работы. Результаты работы по разработке метрики TTED на множестве текстовых деревьев были представлены мной 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 мною результатов по ссылке: `https://github.com/intsystems/Sobolevsky-BS-Thesis`.

Обозначения и сокращения

- *БЯМ* — большая языковая модель (large language model);
- Под *иерархическими сводками*, *картами знаний* или *интеллект-картами* в данной работе будут подразумеваться интеллект-карты на основе предложений (salient-sentence-based mind maps, SSM [6]).
- *KSM* — интеллект-карта на основе главных выдержек (key-snippet based mind map [6]).
- *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).
- *TED* — расстояние редактирования дерева (tree edit distance [8]).

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 текстовая иерархия, то есть дерево, метками вершин которого являются фрагменты текста. Определим этот объект формально. Пусть задан словарь $\mathcal{W}$ и соответствующее множество $\mathcal{S}$ фрагментов текста, составленных из слов этого словаря:

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

Определим текстовое дерево как дерево $T = (V, E)$, $E \subset V^2$, где ля каждой вершины $v \in V$ задана текстовая метка $s(v) \in \mathcal{S}$. 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 предложений из $\mathcal{S}$: $\mathcal{D} = (s_i)_{i=1}^{|\mathcal{D}|}$, $s_i \in \mathcal{S}$, а также иерархическая сводка $T^* \in \mathcal{T}$ данного документа, построенная экспертом. Пусть также задана метрика качества $\mathcal{I}$ генерации текстовых деревьев $T \in \mathcal{T}$ по документам, в общем случае зависящая от справочной карты $\mathcal{T}^*$ и самого документа $\mathcal{D}$, то есть $\mathcal{I} : (T, T^*, \mathcal{D}) \mapsto x \in \mathbb{R}$. Тогда требуется найти отображение $f : \mathcal{D} \mapsto T \in \mathcal{T}$, максимизирующее данную метрику качества $\mathcal{I}$:

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

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 карты, ее смысловое содержание, ранжирование фактов в сводке и др.
- Качество сводки T как краткого представления документа $\mathcal{D}$, например, полнота сводки или фактическое соответствие T и $\mathcal{D}$.
- Качество сводки самой по себе, например, ее неизбыточность, связность 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 оптимизационная задача (1) перепишется в более общем виде как

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

1.3 Задача оптимизации промптинга БЯМ *(TO-DO: обновить)*

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 *(TO-DO: сделать ее отдельным параметром)*, *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 $\mathcal{Q}$.

1.4 Требования к мере близости текстовых деревьев.

Чтобы определить адекватную метрику на множестве текстовых деревьев, мы 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 $\rho : \mathcal{T}^2 \to [0, +\infty)$, отвечающую следующим требованиям учета семантической и структурной близости:

1. Симметричность: $\rho(T, T') = \rho(T', T)$.
2. Равенство нулю в случае равенства аргументов: $\rho(T, T) = 0$.
3. 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 $\rho(T, T') = f(r(v, v'))$.
4. ρ удовлетворяет неравенству треугольника:

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

Обозначим теперь некоторые требование 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 смысловому содержанию), так и по *структуре*. Во-вторых, информативная метрика должна слабо реагировать на несущественные отличия — например, на *парафразирование* предложений в вершинах дерева. Следовательно, среднее значение метрики для деревьев, различающихся по первым двум признакам, должно быть меньше относительно среднего значения расстояния между деревьями, полученными друг из друга парафразированием.

Формализуем эти требования. Обозначим за $P(T)$ множество деревьев, которые можно получить из T парафразированием его меток, $S(T)$ — множество деревьев, составленных из того набора вершин с теми же метками, 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 так: $M(T) = \mathcal{T}_{\sim T} \setminus P(T)$, где $\mathcal{T}_{\sim T}$ — множество деревьев с такой же структурой, как у T. Тогда качественные требования, сформулированные выше, будут отражены формально в следующих соотношениях:

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

Обозначим

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

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

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

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

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

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* (abstractive), 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 *(TO-DO: ссылка на источники терминов)*.

Хотя задача суммаризации появилась в научной литературе ещё во второй половине XX века [9], 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 [10]. Абстрактивная суммаризация начала активно развиваться с момента появления трансформерных архитектур и других архитектур глубокой машинной обработки текста [11].

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 [12]. 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 [13]. 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 [2]. 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. [14], однако впервые она была формализована в работе [4]. 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 [4], подразумевает иерархическую кластеризацию предложений текста с последующей суммаризацией каждого кластера. Целевая функция в [4] 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 предпочти-

телен среди пользователей, чем классические подходы к суммаризации новостных текстов, данная методика не получила дальнейшего развития. Более применимыми стали подходы, основанные на генерации *интеллект-карт* на основе текстов.

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*, поскольку они являются наиболее наглядными и удобными для освоения новой информации 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 [6]) и *интеллект-карты на основе главных выдержек* (key snippet-based mind maps, KSM [6]). 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).

С момента появления в медиа термина «mind map» 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 [1]. Подробный современный обзор применения интеллект-карт в образовании в XXI веке можно найти, например, в [15].

TO-DO: Найти ещё исследований в "многочисленные исследования"

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 [6] 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 [7], 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 [5] 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 [16]. В результате в [5] 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 [3] 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 [3], несколько отличается от рассматриваемого в нашей работе, данное исследование подкрепляет предположение о том, что при достаточно изящной стратегии промптинга БЯМ могут дать качественное решение поставленной нами задачи.

TO-DO: Коммерческие сервисы

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 [17], 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 [18], METEOR [19], MoverScore [20] 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 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 [21] 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 [22], Shepherd [23] и Booookscore [24]. 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 методов оценивания суммаризации с помощью БЯМ можно найти в [12].

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 [21], [25], 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 предложений и детектирования парафраз были достигнуты с использованием моделей на основе трансформерных архитектур; в частности, с использованием моделей на основе энкодера BERT и им подобных [26, 27]. *TO-DO: более подробный обзор*

Сравнение иерархий. *TO-DO: параграф про метрики, с помощью которых сравнивают деревья.*

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. Иерархическая сводка представляет собой дерево из текста, и потому адекватное сравнение таких объектов должно быть многокритериальным и учитывать как семантику предложений в вершинах, так и структуру иерархии.

В имеющихся на данный момент работах по данной теме текстовые деревья, как правило, сравниваются отдельно по своей структуре как деревья и отдельно с помощью, например, метрики ROUGE [17] как наборы текста [5, 6]. 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 [21]. *TO-DO: описать подробнее существующие методы.*

TO-DO:

- Секция обзора про промптинг

3 Предлагаемый метод

3.1 Метрика для сравнения текстовых деревьев

Проанализируем требования к метрике для сравнения текстовых деревьев, сформулированные в разделе 1.4. 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 отмечу, 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 соответствием с выдвинутыми требованиями.

Для вычисления TED мы будем применять алгоритм Zhang-Shasha, а именно его модификацию для неупорядоченных деревьев [28]. 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, то есть степень их сходства по смыслу. На данный момент, однако, не существует ни общепринятой меры семантической близости предложений, ни алгоритма для ее автоматического вычисления, поэтому для практического воплощения предлагаемого метода нам придется оценивать степень семантического сходства. Для этого мы предлагаем использовать в качестве оценки расстояние между эмбеддингами заданных предложений, 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 [27]. В таком случае функцию расстояния естественно определить как $\rho_n(A, B) = 1 - S_C(A, B)$.

Модификации алгоритма Для усовершенствования полученного алгоритма мы предлагаем несколько эвристических модификаций.

1. **Использование контекста.** Зачастую на практике некорректно сравнивать предложения 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 своей реализации мы добавляем возможность при сравнении деревьев предварительно добавлять в метки вершин все предложения из родительских вершин в качестве контекста перед предложением в вершине и после сравнивать эмбеддинги, полученные с помощью модели с учетом этого контекста.

2. **Фактор глубины.** В зависимости от приложения предложенного алгоритма различие предложений в листьях дерева и в вершинах, близких к корневой, может считаться более или менее значимым. В связи с этим мы определяем коэф-

фициент глубины γ, позволяющий изменять вес степени сходства между предложениями в вершинах в зависимости от их глубины: $f(r(v, v')) = \gamma^d r(v, v')$, где d — глубина вершины v.

3. **Предварительное вычисление.** Многократное вычисление эмбеддингов с помощью нейросетевой модели может быть очень затратно по времени для больших деревьев, поэтому мы предлагаем предварительно вычислять эмбеддинги для всех предложений в вершинах и применять предложенный алгоритм уже для дерева из эмбеддингов с вышеуказанной стоимостью обновления меток.

Базовый метод. Для сравнения с предложенной метрикой я использую функцию сходства текстовых деревьев, использованную в работе [5] для оценки сходства автоматически сгенерированных интеллект-карт 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', $\mathrm{ROUGE}(v, v')$ — усредненная оценка ROUGE-1, ROUGE-2 и ROUGE-L сходства $s(v)$ и $s(v')$:

$$\mathrm{ROUGE}(e, e') = \frac{1}{3}\left(\mathrm{ROUGE\text{-}1}(e_1, e'_1) + \mathrm{ROUGE\text{-}2}(e_1, e'_1) + \mathrm{ROUGE\text{-}L}(e_1, e'_1)\right).$$

Поскольку предлагаемый нами метод оценивает расстояние между текстовыми аргументами, а baseline-метод оценивает их сходство, то дл единообразия в качестве функции расстояния для сравнения мы будем использовать $\rho_0(T, T') = \mathrm{Sim}^{\max} - \mathrm{Sim}(T, T')$, где $\mathrm{Sim}^{\max}$ — максимально возможное значение близости для дерева заданного размера, равное близости идентичных деревьев: $\mathrm{Sim}^{\max} = \mathrm{Sim}(T, T)$.

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 сэмплирования 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 второго 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 $R_S(\rho)$. ∎

Имея выборку $\mathcal{D}$ обозначенного выше вида и несмещенные оценки $R_S(\rho)$ и $R_M(\rho)$ 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 интеллект-карт

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 интеллект-карт с авторскими по каждому аспекту сходства отдельно. В связи с этим воспользуемся несколькими метриками сходства интеллект-карт, каждая из которых будет отражать аспекты сходства своей природы. Выделим три основных аспекта, по которым будем сравнивать текстовые деревья отдельно:

1. *Семантика* дерева, то есть его смысловое содержание в отрыве от древовидной структуры;
2. *Структура* дерева без учета текстовых меток вершин;
3. *Ранжирование* фактов в дереве, то есть распределение предложений в вершинах по уровням иерархии.

Семантическое сходство. Первый измеряемый в данной работе аспект сходства предполагает сходство 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 предложений. Применим одну из таких метрик — *BERTScore* [22], основанную на сравнении предложений по их представлениям, полученным с помощью BERT-подобной языковой модели.

Структурное сходство. Второй аспект, по которому можно сравнивать интеллект-карты — их структура. Для сравнения структур текстовых деревьев можно абстрагироваться от меток их вершин и сравнивать их как неразмеченные деревья. Существуют разные метрики сравнения иерархий, однако остановимся на метрике, которая уже была применена в данной работе и которая нашла широкое применение для сравнения деревьев — *расстоянии редактирования*. В данном случае способом сравнить исключительно структуры деревьев будет применить алгоритм поиска расстояния редактирования для них, заменив все метки на одинаковые и 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.

Сходство ранжирования. Третий выделяемый в данном разделе аспект сходства иерархий — сходство в ранжировании предложений в вершинах по уровням иерархии. В терминах оценки качества иерархической суммаризации это означает качество расположения фактов, выносимых в иерархическую сводку, по порядку от главного к деталям. Сразу применить для оценки ранжирования стандартные метрики качества здесь мешают две особенности задачи. Во-первых, предложения в сгенерированной и авторской сводках почти наверное будут отличаться формулировками, хотя могут при этом совсем не отличаться по смыслу. Во-вторых, следует учесть, что на одном уровне иерархии в дереве могут находиться несколько вершин, поэтому нужно учесть нахождение на одном месте в рейтинге сразу многих вершин.

Первую проблему можно решить, сопоставив попарно по семантической близости предложения в вершинах двух деревьев и найдя в них пары ближайших по смыслу предложений, а затем оценивая ранжирование предложений по экспертному ранжированию их пар. Здесь возникает очевидный вопрос: насколько семантически близкими должны быть предложения, чтобы их можно было сопоставить друг другу как один и тот же факт? Для получения четкого ответа на этот вопрос пришлось бы обучать границу отсечения на выборке для задачи детекции парафразов, однако для целей данной работы можно воспользоваться значениями, уже подобранными в работе [27] для используемых языковых моделей.

Для сравнения ранжирования тех предложений во втором дереве, для которых мы нашли соответствующие предложения в первом, можно воспользоваться *коэффициентом ранговой корреляции Спирмена* [29]. При этом предложения, которым не будет найдено соответствие в другом дереве, не будем учитывать при подсчете, поскольку полноту иерархической сводки уже будет оценивать метрика семантического сходства.

TO-DO: как минимум, отдельный раздел про то, как мы будем пытаться последовательными промптами заставить БЯМ генерировать карты

4 Вычислительные эксперименты

4.1 Тестирование метрики для сравнения текстовых деревьев

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

1. текстового дерева T;
2. деревьев, которые идентичны по семантическому значению и структуре, но предложения в узлах дерева *перефразированы* — подвыборка $\mathcal{T}_1$ из $P(T)$;
3. деревьев, которые сформированы из одних и тех же предложений, но с разной *структурой* дерева — подвыборка $\mathcal{T}_2$ из $S(T)$;
4. деревьев, которые идентичны по структуре и схожи по наборам слов в предложениях, 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). Введу 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 соответствующего значения $\overline{\rho}_1$ от соответствующих значений $\overline{\rho}_2$ и $\overline{\rho}_3$.

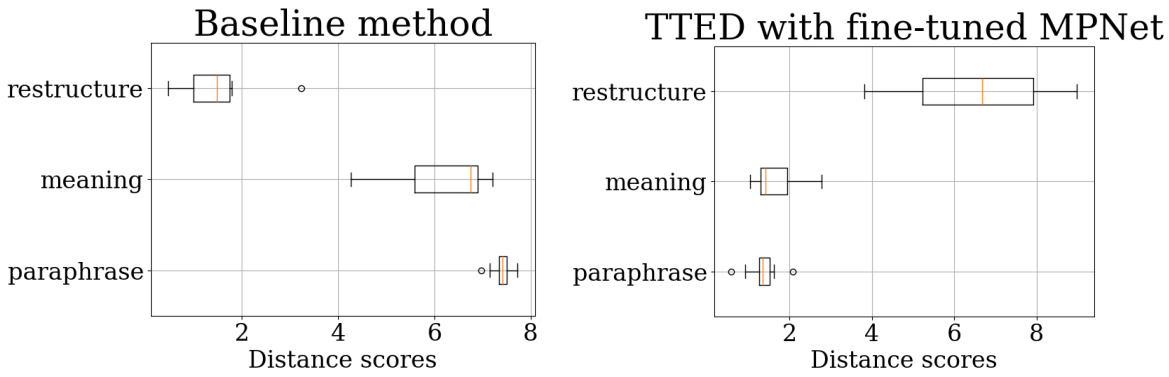
Экспериментальные данные. Поскольку привлечь других экспертов или краудсорсинг к созданию выборки для данной работы не представилось возможным, а ручное создание интеллект-карт занимает довольно много времени, для создания выборки для тестирование метрик на чувствительность к различным аспектам сходства текстовых деревьев был привлечен генеративный ИИ. Процесс генерации данных состоял из следующих этапов:

1. Создание базовой интеллект-карты по научному исследованию вручную.
2. Генерация парафразов, реструктуризаций и переосмыслений этой интеллект-карты при помощи запросов к БЯМ DeepSeek V3 (TO-DO: ссылка на модель).
3. Ручная проверка сгенерированных данных на соответствие запросу.

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.

Результаты. Результаты тестирования различных языковых моделей с расстоянием на основе косинусного коэффициента в сравнении с baseline-методом представлены в таблице 1. Оценки расстояния, полученные с помощью baseline-алгоритма и нашего алгоритма с использованием модели `paraphrase-multilingual-mpnet-base-v2` из библиотеки `sentence_transformers` представлены на рис. 1a и 1b соответственно.

На примере версии TTED с языковой моделью MPNet были проведены эксперименты с разными эмбеддинговыми расстояниями и без использования контекста при вычислении эмбеддингов. Результаты этих экспериментов представлены в таблице 2.



(a) Оценки базового метода (b) Оценки TTED

Рис. 1: Оценки расстояний с помощью TTED и baseline-метода

Модель	$\overline{\rho}_1$	$\overline{\rho}_2$	$\overline{\rho}_3$	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Baseline	7,41±0,22	1,50±0,71	6,25±0,96	6,29±3,58	1,22±0,24
DistilRoBERTa	2,38±0,35	6,68±1,66	2,83±0,59	0,38±0,11	0,89±0,28
SPECTER	0,76±0,21	1,96±0,56	0,91±0,34	0,41±0,14	0,92±0,27
MPNet	1,72±0,60	6,33±1,76	1,87±0,74	0,27±0,07	1,07±0,59
Paraphrase MPNet	1,35±0,38	6,51±1,70	1,68±0,56	**0,21±0,05**	**0,88±0,37**

Таблица 1: Средние оценки расстояния с помощью разных языковых моделей

Видно, что TTED в среднем показывает значения расстояния больше для деревьев, различающихся по семантике и структуре, чем для деревьев, которые являются перефразированием друг друга. По рис. 1b и значениям $R_2(\rho)$ в таблице 1 видно, что TTED сильнее всего отражает различия деревьев по своей структуре, однако в ходе экспериментов удалось подобрать такую языковую модель *(TO-DO: ссылка на модель)* для построения эмбеддингов, с использованием которой разница между $\overline{\rho}_1$ и $\overline{\rho}_3$ значительна (области значений расстояний на $\mathcal{T}_1$ и $\mathcal{T}_3$ различимы в пределах своих среднеквадратичных отклонений). Это подкрепляет утверждение о том, что TTED отражает значимые отличия текстовых деревьев заметно сильнее, чем незначительные.

Для сравнения рассмотрим значения, полученные с помощью базового метода (рис. 1a, таблица 1). Мы видим картину, практически противоположную полученной с помощью TTED: наибольшие средние значения расстояний получены между деревьями-парафразами, в то время как расстояния между деревьями с разной структурой, хоть и имея довольно большой разброс, в среднем меньше остальных. Это связано с тем, что в основании базового метода лежит метрика ROUGE, оперирующая на уровне слов и словосочетаний, но не учитывающая реальную семантику предложений, из-за чего более неочевидные смысловые различия не так увеличивают значение расстояния, как перефразирование с изменением большей части слов пред-

ложений. Данный эксперимент показывает, что базовый метод сравнения текстовых деревьев отражает значимые отличия деревьев не лучше, чем незначительные.

$r(x, y)$	$\overline{\rho}_1$	$\overline{\rho}_3$	$R_M^{\mathcal{D}}(\rho)$
$1 - S_C(x, y)$	1,35±0,38	1,68±0,56	**0,88±0,37**
$\|\|x - y\|\|_2$	13,79±2,14	15,17±2,45	0,93±0,20
$\|\|x - y\|\|_1$	291,28±44,43	322,24±51,94	0,92±0,19

Таблица 2: Средния значения TTED для разных эмбеддинговых расстояний

TO-DO: поставить новый эксперимент по вычислению времени работы

Заключение

TO-DO:

Обсуждение результатов.

Ограничения исследования.

Направления дальнейшей работы.

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 используемые до этого методы.
- *(в процессе)* Проведено многокритериальное исследование нескольких методов иерархической суммаризации при помощи больших языковых моделей с использованием предложенных новых методов сравнения текстовых деревьев как методов сравнения сгенерированных текстовых иерархий с экспертными.

Список литературы

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

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

[29] Spearman C. The Proof and Measurement of Association between Two Things // The American Journal of Psychology. — 1904. — Vol. 15, no. 1. — P. 72–101. — Access mode: `http://www.jstor.org/stable/1412159` (online; accessed: 2025-06-13).