Применение больших языковых моделей для иерархической суммаризации текстов научных публикаций

Соболевский Ф. А.
Научный руководитель: д. ф.-м. н. Воронцов К. В.

Московский физико-технический институт

2025

Цели исследования

- Применить большие языковые модели (БЯМ) для иерархической суммаризации научных статей и определить оптимальный метод работы с моделью, позволяющий максимизировать качество генерации для выбранной БЯМ.
- Предложить новый способ измерения качества иерархической суммаризации, основанный на многокритериальном сравнении текстовых деревьев.
- Формализовать требования к адекватности метрики на множестве текстовых деревьев и исследовать свойства предложенной метрики.

Основная идея

- **Объект генерации** — *текстовые деревья*.
- **Критерий качества** — сходство с авторской сводкой.
- **Проблема:** как сравнивать текстовые деревья?

- **Многокритериальная** оценка сходства (по структуре, семантике и т. д.);
- **TTED** — информативная единая метрика на множестве текстовых деревьев;
- **Качество метрики** — чувствительность к *значимым* различиям по отношению к *незначимым*.

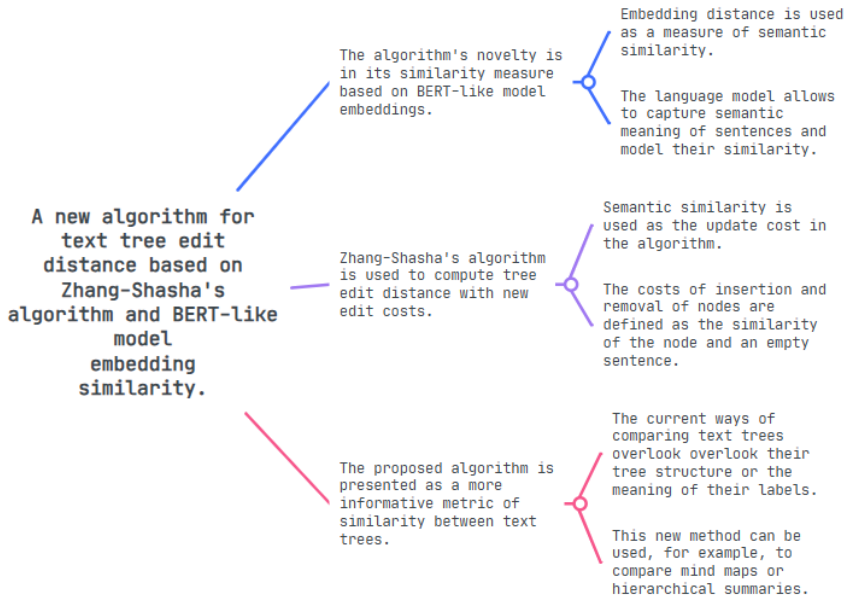

Рис.: Пример иерархической сводки по научному исследованию

Постановка задачи иерархической суммаризации

- Пусть $\mathcal{S}$ — множество возможных фрагментов текста.
- Текстовое дерево — дерево $T = (V, E)$, где $E \subset V^2$ и для каждого $v \in V$ определен текст $s(v) \in \mathcal{S}$.
- $\mathcal{T}$ — множество рассматриваемых текстовых деревьев.
- $\rho : T^2 \to \mathbb{R}^+$ — метрика на множестве текстовых деревьев.
- Задача: найти отображение $f : D \mapsto T$, строящее иерархическую сводку $T \in \mathcal{T}$ по документу D, минимизирующее ее отличие от авторской сводки T^*:

$$\rho(f(D), T^*) \longrightarrow \min_f .$$

Требования к метрике на множестве текстовых деревьев

Пусть $T, T' \in \mathcal{T}$. Зададим следующие требования к метрике ρ на множестве $\mathcal{T}$:

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

$$\forall T, T', T'' \in \mathcal{T} \quad \rho(T, T'') \leq \rho(T, T') + \rho(T', T''). \qquad (1)$$

4. Существует некоторая неубывающая функция $f : \mathbb{R}^+ \to \mathbb{R}^+$, такая что:
   - 4.1 Если T' получено из T добавлением в T вершины v, то $\rho(T, T') = f(r(v))$;
   - 4.2 Если T' получено из T удалением из T вершины v, то $\rho(T, T') = f(r(v))$;
   - 4.3 Если T' получено из T заменой вершины v на v', то $\rho(T, T') = f(r(v, v'))$.

Предлагаемая метрика — *TTED*

- ***TTED*** *(text tree edit distance)* — расстояние редактирования[1], стоимости операций редактирования в котором определяются заданной мерой семантического расстояния между текстами в вершинах.
- В качестве метрики семантического расстояния можно применить языковую модель $\mathrm{LM} : S \to \mathbb{R}^n$ и 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 $\mathbb{R}^n$.

Используемые эвристики:

- Использование родительских вершин в качестве **контекста** для текстов в дочерних.
- **Предварительное вычисление** эмбеддингов и попарных расстояний для всех текстов в вершинах.

[1] *Zhang Kaizhong, Statman Richard, Shasha Dennis.* On the Editing Distance Between Unordered Labeled Trees

Базовый метод

- Для сравнения используется оценка сходства текстовых деревьев из работы *Zhang et al., 2024*[2]. Для текстовых деревьев $T = (V, E)$ и $T' = (V', E')$ она определяется как:

$$\mathrm{Sim}(T, T') = \min_{P \subset E \times E'} \sum_{(e,e') \in P} \sum_{i=0,1} \mathrm{ROUGE}(e_i, e'_i).$$

  где P — однозначное сопоставление ребер T ребрам T' (оптимальное ищется жадным алгоритмом), $\mathrm{ROUGE}(v, v')$ — усредненная оценка ROUGE-1, ROUGE-2 и ROUGE-L сходства $s(v)$ и $s(v')$.
- В экспериментах для единообразия в качестве оценки расстояния используется $\rho(T, T') = \mathrm{Sim}^{\max} - \mathrm{Sim}(T, T')$, где $\mathrm{Sim}^{\max} = \mathrm{Sim}(T, T)$.

[2] *Zhang Zhuowei, Hu Mengting, Bai Yinhao, and Zhang Zhen.* Coreference Graph Guidance for Mind-Map Generation

Критерии качества метрики

Пусть для $T \in \mathcal{T}$:

- $P(T)$ — множество деревьев-парафразов T;
- $S(T)$ — множество деревьев-реструктуризаций T;
- $M(T)$ — набор деревьев с такой же структурой, как у T, но с разной семантикой: $M(T) = \mathcal{T}_{\sim T} \setminus P(T)$.

Задача оптимизации:

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

где

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

$$r_S(\rho, T) = \mathbb{E}_{T' \sim P(T),\ T'' \sim S(T)} \left[\frac{\rho(T, T')}{\rho(T, T'')} \right],$$

$$r_M(\rho, T) = \mathbb{E}_{T' \sim P(T),\ T''' \sim M(T)} \left[\frac{\rho(T, T')}{\rho(T, T''')} \right].$$

Оценка коэффициентов качества по выборке

Рассмотрим случайную выборку текстовых деревьев $\mathcal{D} = \{T, T'_1, \ldots, T'_p, T''_1, \ldots, T''_s, T'''_1, \ldots, T'''_m\}$, 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 $\mathcal{D}$:

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

Теорема (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 случайной выборке $\mathcal{D}$:

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

Многокритериальное сравнение текстовых деревьев

Для сравнения текстовых деревьев по различным аспектам сходства применимы следующие метрики:

- **Семантическое сходство:** сравнение текстов из вершин деревьев как линейных с помощью *BERTScore*[3];
- **Структурные различия:** сравнение деревьев без разметки с помощью *расстояния редактирования (TED)*;
- **Сходство ранжирования** предложений в иерархии: сопоставление предложений в вершинах по семантической близости и сравнение ранжирования с помощью *коэффициента корреляции Спирмена*[4].

---

[3] *Zhang Tianyi, Kishore Varsha, Wu Felix, Weinberger Kilian Q, Artzi Yoav.* BERTScore: Evaluating text generation with BERT

[4] *Spearman Charles.* The Proof and Measurement of Association between Two Things

Тестирование метрик — постановка эксперимента

Для оценки семантической близости в TTED использовался ряд языковых моделей из библиотеки `sentence-transformers`[5].
Эксперименты — вычисление расстояний на выборке, состоящей из:

1. Основного дерева T, с которым сравнивались остальные;
2. Парафразов T (подвыборка `paraphrase`);
3. Реструктуризаций T (подвыборка `restructure`);
4. Деревьев, отличных от T только по смыслу (подвыборка `meaning`).

Цель эксперимента — найти среди предложенных такую метрику ρ, для которой будут минимальными оценки $R_S^{\mathcal{D}}(\rho)$ и $R_M^{\mathcal{D}}(\rho)$.

[5] `https://sbert.net/`

Тестирование метрик — результаты

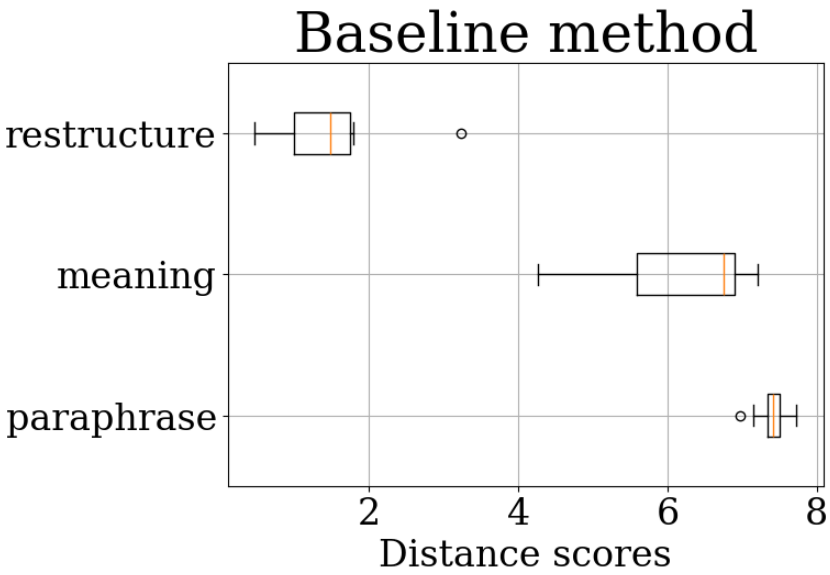

Рис.: Оценки базового метода

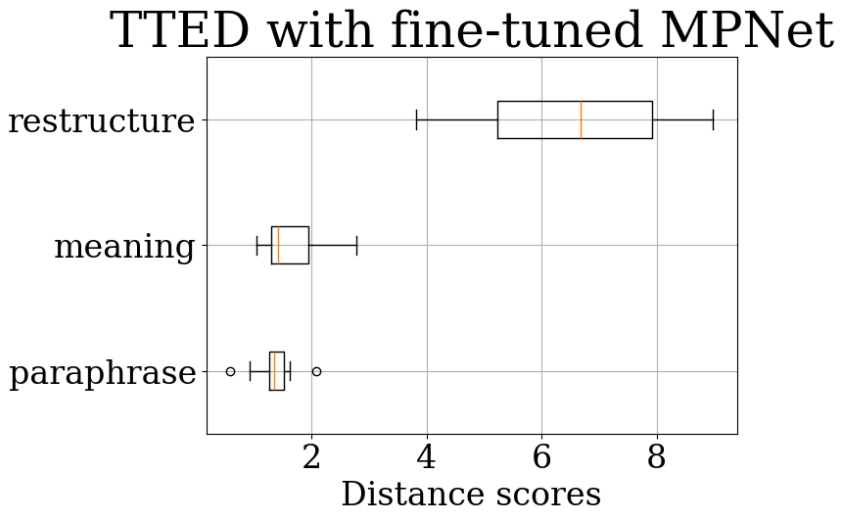

Рис.: Оценки нашего метода

Модель	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Baseline	6,29±3,58	1,22±0,24
DistilRoBERTa	0,38±0,11	0,89±0,28
SPECTER	0,41±0,14	0,92±0,27
MPNet	0,27±0,07	1,07±0,59
Fine-tuned MPNet	**0,21±0,05**	**0,88±0,37**

Тестирование БЯМ — постановка эксперимента

Используемая для тестирования БЯМ — модель `mistral-large-latest` из библиотеки `langchain-mistralai`[6].
Исследованные методы генерации иерархических сводок:

1. **Прямой промптинг** (direct prompting) модели созданными вручную запросами;
2. **Оптимизация запросов** (prompt optimization) с помощью библиотеки `langmem`[7] с использованием многокритериальной оценки генерации с помощью человеческих запросов;
3. **Последовательный промптинг** (sequential prompting) модели с выбором пользователем генерируемых вершин дерева.

Метрики сходства по различным аспектам — BERTScore, TED, r_s, а также единая метрика TTED.

[6] `https://python.langchain.com/docs/integrations/providers/mistralai/`
[7] `https://langchain-ai.github.io/langmem/`

Тестирование БЯМ — результаты

Метод	BERTScore	TED	TTED
Direct prompting	0,41	11,0	14,13
Prompt optimization	0,72	4,0	7,26
Sequential prompting	**0,78***	**0,0***	**3,35***

Таблица: Результаты тестирования БЯМ для иерархической суммаризации

**При оптимальном сценарии взаимодействия пользователя с системой.*

Заключение

Положения, выносимые на защиту

- Введен новый коэффициент качества метрик на множестве текстовых деревьев и предложена несмещенная оценка данного коэффициента по случайной выборке текстовых деревьев.
- Разработан новый алгоритм оценки расстояния между текстовыми деревьями, лучше отражающий значимые отличия текстовых деревьев в терминах введенного коэффициента качества, чем используемые до этого методы.
- Проведено многокритериальное исследование методов иерархической суммаризации при помощи БЯМ с использованием предложенных новых методов сравнения с экспертными сводками.

Литература

- *Zhang Z., Hu M., et al.* Coreference Graph Guidance for Mind-Map Generation // Proceedings of the AAAI Conference on Artificial Intelligence. — 2024. — Vol. 38. — P. 19623–19631.
- *Zhang K., Statman R., Shasha D.* On the editing distance between unordered labeled trees. // Information processing letters. 1992 May 25; 42(3): 133-9.
- *Vrbanec T., Meštrović A.* Comparison study of unsupervised paraphrase detection: Deep learning — The key for semantic similarity detection. // Expert systems. 2023 Nov; 40(9): e13386.