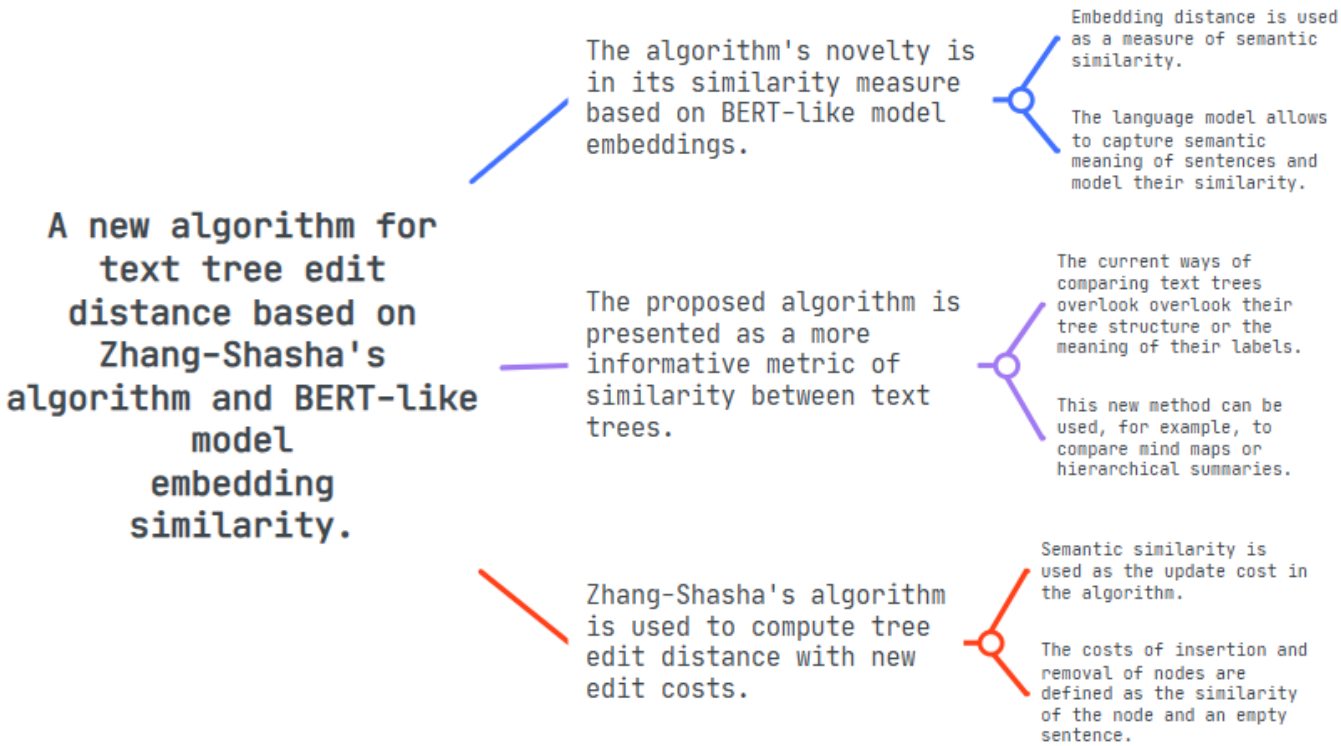
Расстояние редактирования текстового дерева: сравнение текстовых иерархий с использованием языковых моделей

Соболевский Ф. А.[1], д. ф.-м. н. Воронцов К. В.[1,2]

[1]Московский физико-технический институт
[2]Московский государственный университет им. М. В. Ломоносова

2025

Мотивация исследования



Пример текстового дерева — иерархическая сводка данного исследования в виде *интеллект-карты*

Проблема: как сравнивать иерархические сводки между собой, учитывая как их структуру, так и семантику?

Постановка задачи иерархической суммаризации

Пусть $\mathcal{S}$ — *множество текстов* над заданным словарем.
Текстовое дерево — дерево $T = (V, E)$, где $E \subset V^2$ и для каждого $v \in V$ определен текст $s(v) \in \mathcal{S}$.
$\mathcal{T}$ — рассматриваемое *множество текстовых деревьев*.

Задача: найти отображение $f : D \mapsto T$, строящее иерархическую сводку $T \in \mathcal{T}$ по документу D, минимально отличающуюся от эталонной сводки T^*, построенной экспертом:

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

Вопрос: как выбрать метрику $\rho : \mathcal{T}^2 \to \mathbb{R}_+$?

Требования к метрике на множестве текстовых деревьев

Пусть задана функция **семантического (смыслового) расстояния** между текстами: $r : \mathcal{S}^2 \to [0, +\infty)$.
Обозначение: $r(v, v') := r(s(v), s(v'))$, а $r(v) := r(s(v), \lambda)$, где λ — пустая строка.

Требования к метрике $\rho : \mathcal{T}^2 \to \mathbb{R}_+$:

1. ρ является корректной метрикой на $\mathcal{T}$.
2. Пусть $T, T' \in \mathcal{T}$. Существует некоторая неубывающая функция $f : [0, +\infty) \to [0, +\infty)$, такая что:
   - 2.1 Если T' получено из T добавлением в T вершины v, то $\rho(T, T') = f(r(v))$;
   - 2.2 Если T' получено из T удалением из T вершины v, то $\rho(T, T') = f(r(v))$;
   - 2.3 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], где стоимость операций редактирования:
а) *замены* вершины v на v': $r(s(v), s(v'))$;
б) *добавления/удаления* вершины v: $r(s(v), \lambda)$;
где λ — пустая строка.

Семантическое расстояние r можно измерять как расстояние между *эмбеддингами* (векторными представлениями) текстов, полученными с помощью языковой модели $\mathrm{LM} : \mathcal{S} \to \mathbb{R}^n$:

$$\forall s, s' \in \mathcal{S} \quad r(s, s') = \rho_n(\mathrm{LM}(s), \mathrm{LM}(s')),$$

где ρ_n — метрика в $\mathbb{R}^n$.

[1] *Zhang Kaizhong, Statman Richard, Shasha Dennis.* On the Editing Distance Between Unordered Labeled Trees (1992)

Базовый метод сравнения текстовых деревьев

В работе *Zhang et al., 2024*[2] близость текстовых деревьев $T = (V, E)$ и $T' = (V', E')$ определяется как

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

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 $s(v')$.

Для единообразия в качестве оценки расстояния используется

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

[2] *Zhang Zhuowei, Hu Mengting, Bai Yinhao, and Zhang Zhen.* Coreference Graph Guidance for Mind-Map Generation (2024)

Аспекты различия текстовых деревьев

Пусть для $T \in \mathcal{T}$ заданы множества деревьев:

1. $P(T)$ — отличающихся от T только перефразированием;
2. $S(T)$ — отличающихся от T только структурой;
3. $M(T)$ — отличающихся от T только семантикой (по смысловому содержанию).

Идея: для адекватной метрики ρ на $\mathcal{T}$ должно выполняться

$$\langle \rho(T, T') \rangle_{T' \in P(T)} \ll \langle \rho(T, T'') \rangle_{T'' \in S(T)},$$

$$\langle \rho(T, T') \rangle_{T' \in P(T)} \ll \langle \rho(T, T''') \rangle_{T''' \in M(T)}.$$

Критерии качества метрики

Рассмотрим выборку
$\mathcal{D} = \{T, T'_1, \ldots, T'_p, T''_1, \ldots, T''_s, T'''_1, \ldots, T'''_m\}$, где $T \in \mathcal{T}$,
$T'_i \in P(T)$, $T''_j \in S(T)$, $T'''_k \in M(T)$.
Коэффициенты качества метрики ρ по выборке $\mathcal{D}$:

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

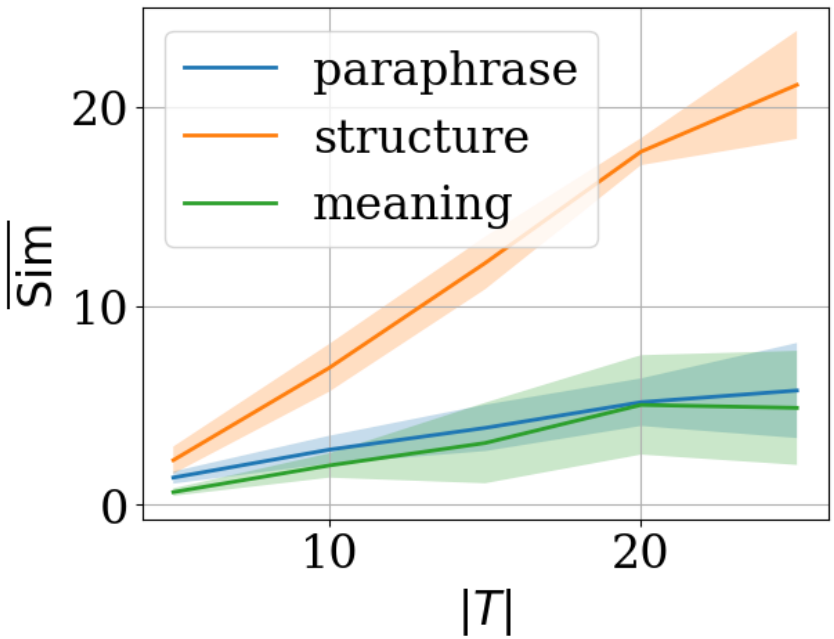
$R_S^{\mathcal{D}}(\rho)$ — чувствительность метрики ρ к **парафразированию** по отношению к **структуре**;
$R_M^{\mathcal{D}}(\rho)$ — чувствительность к **парафразированию** по отношению к **семантике**.
Оптимизационная задача:

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

Тестирование метрик — результаты

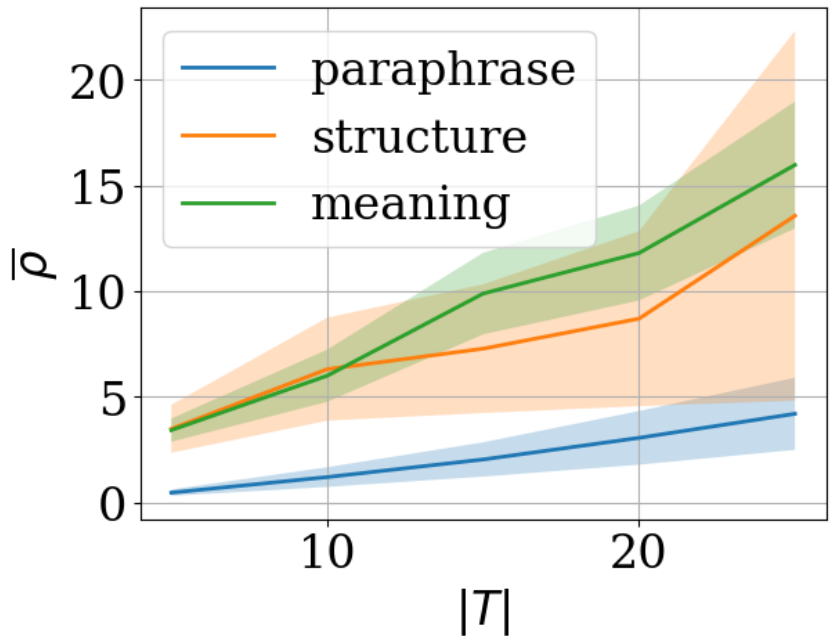
Зависимость от размера дерева $|T|$:

а) средних значений базового коэффициента сходства



$\text{Sim}(\cdot, \cdot)$ сходно отражает различия по семантике и парафразированию, заметно меньше — по структуре.

б) средних расстояний по метрике TTED



TTED отражает различия по парафразированию заметно меньше, чем по структуре и семантике.

Тестирование метрик — результаты

Результаты тестирования базового метода и TTED с разными моделями-кодировщиками для получения эмбеддингов текстов на синтетических данных

Модель	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Базовый метод	1,44±0,25	0,89±0,03
TTED с DistilRoBERTa	0,54±0,13	0,61±0,10
TTED с SPECTER	0,45±0,10	0,41±0,06
TTED с MPNet	0,40±0,11	0,42±0,08
TTED с дообученной MPNet	**0,38±0,16**	**0,36±0,04**

Значимые отличия в сравнении с незначительными отражаются TTED лучше, чем базовым методом.

Тестирование модификаций TTED

Зависимость коэффициентов качества от **метрики для сравнения эмбеддингов** в TTED

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

Зависимость коэффициентов качества от использования **контекста** в TTED

Метод	$\overline{\rho}_1$	$\overline{\rho}_2$	$\overline{\rho}_3$	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Без контекста	1,25	4,80	4,34	0,32	0,29
С контекстом	1,82	7,71	7,56	**0,24**	**0,24**

Оптимальной и наиболее интерпретируемой конфигурацией TTED является конфигурация с кодировщиком MPNet, расстоянием на основе косинусного коэффициента и с использованием контекста в виде родительских вершин.

Основные результаты

1. Предложена новая метрика на множестве текстовых деревьев — TTED.
2. Показано, что TTED лучше отражает значимые различия текстовых деревьев, чем использованный до этого коэффициент сходства.
3. Подобрана оптимальная конфигурация TTED.
4. Предложенную метрику можно использовать для оценки качества в задачах иерархической суммаризации, построения интеллект-карт и других задачах автоматической генерации текстовых иерархий.

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.