Метод оценки сходства текстовых деревьев с помощью расстояния редактирования и языковых моделей.

Соболевский Ф. А.,
д. ф.-м. н. Воронцов К. В.

Московский физико-технический институт

2025

Цели исследования

- Предложить агрегированный критерий сходства текстовых деревьев по нескольким аспектам сходства.
- Исследовать теоретические свойства предложенной метрики.
- Сравнить предложенный метод с существующими методами оценки сходства текстовых деревьев.

Пример применения: оценивание качества иерархической суммаризации текстов путём сравнения с суммаризациями, построенными экспертами.

Основная идея

- Объект сравнения — **текстовые деревья** (деревья с текстовыми метками вершин).
- Несколько аспектов сходства, в т. ч. **структурные** и **семантические**.
- *Проблема:* существующие критерии сходства текстовых деревьев недостаточно отражают многокритериальность их сходства.
- *Решение* — совместить расстояние редактирования и языковые модели для агрегирования структурных и семантических аспектов сходства.

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.

Постановка задачи

- Пусть $\mathcal{S}$ — множество всевозможных фрагментов текста. Определим текстовое дерево как дерево $T = (V, E)$, где $E \subset V^2$ и для каждого $v \in V$ определена текстовая метка $s(v) \in \mathcal{S}$. Обозначим множество всевозможных текстовых деревьев за $\mathcal{T}$.
- Зададим функцию семантической близости текстовых фрагментов: $r : \mathcal{S}^2 \to [0, +\infty)$. Для $v, v' \in V$ обозначим $r(v, v') := r(s(v), s(v'))$, а $r(v) := r(s(v), \lambda)$, где λ — пустая строка.
- *Найти:* функцию сходства $\rho : \mathcal{T}^2 \to [0, +\infty)$, отвечающую заданным требованиям учета семантической и структурной близости.

Постановка задачи: продолжение

Пусть $T, T' \in \mathcal{T}$. Зададим следующие требования к метрике ρ на множестве $\mathcal{T}$ (здесь f — некоторая неубывающая функция):

1. $\rho(T, T') = \rho(T', T)$.
2. $\rho(T, T') \geq 0$, $\rho(T, T) = 0$.
3. Если T' получено из T добавлением в T вершины v, то $\rho(T, T') = f(r(v))$.
4. Если T' получено из T удалением из T вершины v, то $\rho(T, T') = f(r(v))$.
5. Если T' получено из T заменой вершины v на v', то $\rho(T, T') = f(r(v, v'))$.
6. $\forall T, T', T'' \in \mathcal{T} \quad \rho(T, T'') \leq \rho(T, T') + \rho(T', T'')$.

Из последнего условия естественным образом следует, что расстояние ρ будет соответствовать наименьшему по стоимости набору операций редактирования дерева.

Предлагаемое решение

- Общепринятый критерий структурного сходства деревьев — **расстояние редактирования** (наименьшая стоимость преобразования одного дерева в другое при помощи операций добавления, удаления и замены вершины при заданной стоимости данных операций).
- Требованиям выше удовлетворяет расстояние редактирования, для которого стоимости операций редактирования — $f(r(v))$ и $f(r(v, v'))$ соответственно.
- Для аппроксимации 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 близости в $\mathbb{R}^n$.

Эвристики для модификации алгоритма

- Конкатенация меток вершин с предложениями из родительских вершин в качестве **контекста** предложений в дочерних вершинах.
- Умножение весов операций на **фактор глубины** γ: $f(r(v, v')) = \gamma^d r(v, v')$, где d — глубина вершины v.
- **Предварительное вычисление** эмбеддингов и попарных расстояний для всех предложений в вершинах сравниваемых деревьев. Тогда совершается $O(|V|)$ вызовов языковой модели и $O(|V|^2)$ вызовов функции сходства.

Baseline-метод

- Для сравнения мы используем критерий оценки сходства текстовых деревьев, использованный в работе *Zhang et al., 2024* для оценки сходства автоматически сгенерированных интеллект-карт из предложений с эталонными. Для текстовых деревьев $T = (V, E)$ и $T' = (V', E')$ он определяется как:

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

  где P — однозначное сопоставление ребер T ребрам T', $\text{ROUGE}(v, v')$ — усредненная оценка ROUGE-1, ROUGE-2 и ROUGE-L сходства $s(v)$ и $s(v')$.
- В экспериментах для единообразия мы используем в качестве оценки расстояния $\rho(T, T') = -\text{Sim}(T, T')$.

Вычислительный эксперимент — постановка

Использовалась `distiluse-base-multilingual-cased-v1` из библиотеки `sentence-transformers`.
Эксперименты — вычисление оценок попарного сходства на трех выборках по 5 деревьев:

1. которые идентичны по семантическому значению и структуре, но предложения в узлах дерева *перефразированы* (выборка D_1 — `paraphrase`);
2. которые сформированы из одних и тех же предложений, но с разной *структурой* дерева (выборка D_2 — `restructure`);
3. которые идентичны по структуре и схожи по наборам слов в предложениях, но значительно *отличаются по значению* (выборка D_3 — `meaning`).

Цель алгоритма сравнения деревьев $\rho(\cdot, \cdot)$ — минимизация значений в $\{\rho(T, T')\}_{T,T' \in D_1}$ относительно значений в $\{\rho(T, T')\}_{T,T' \in D_2}$ и $\{\rho(T, T')\}_{T,T' \in D_3}$.

Вычислительный эксперимент — результаты

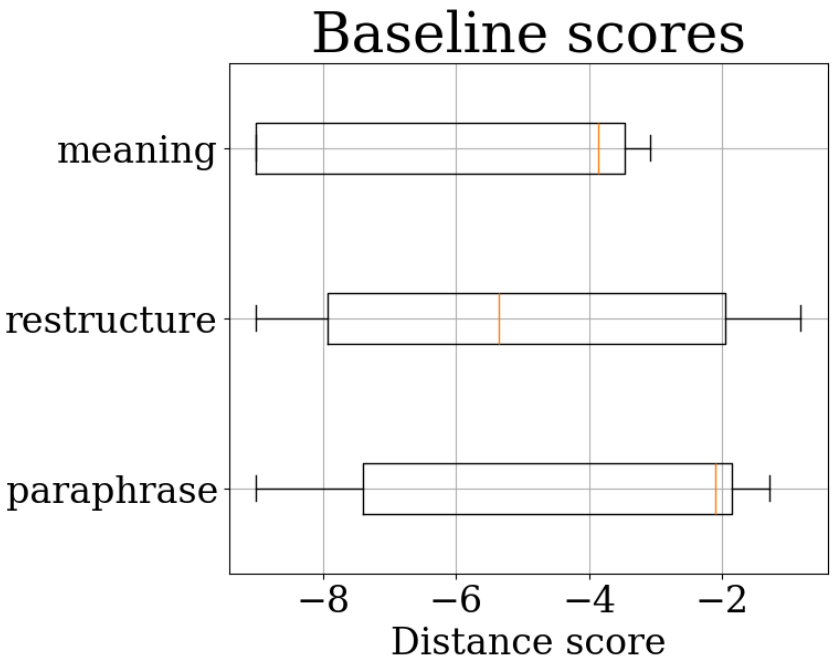

Рис.: Оценки baseline-метода

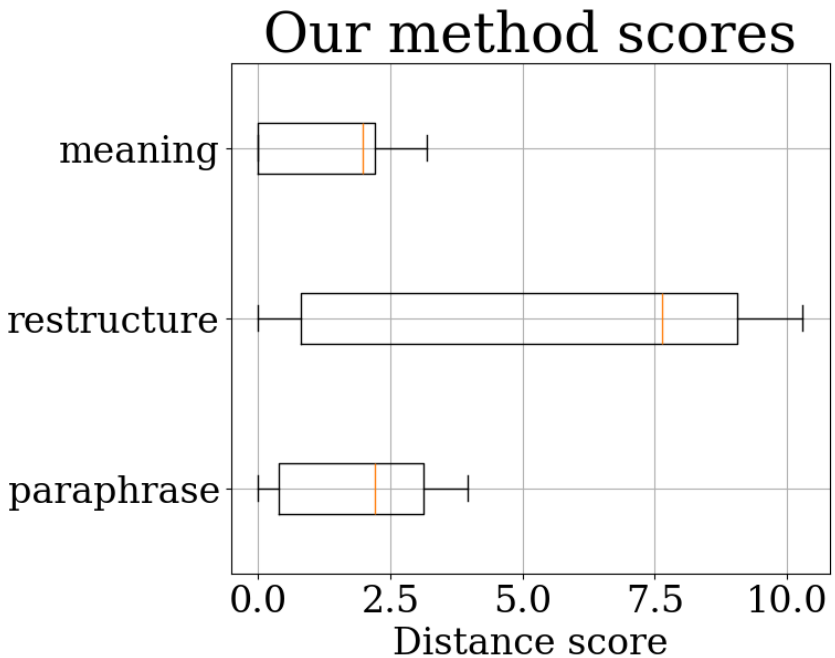

Рис.: Оценки нашего метода

Baseline-метод дает самую высокую среднюю оценку расстояния парам деревьев из выборки `paraphrase` и высокий разброс значений. Наш метод дает самые высокие оценки расстояния в среднем парам деревьев в `restructure` и примерно равные в `meaning` и `paraphrase`.

Вычислительный эксперимент — результаты

Эксперимент	Значения
Meaning	**-5,55 ± 2,60**
Restructure	**-5,14 ± 3,00**
Paraphrase	-3,95 ± 3,21

Таблица: Оценки baseline-метода

Эксперимент	Значения
Meaning	**1,43 ±1,11**
Restructure	$5,75 \pm 4,02$
Paraphrase	**1,95 ± 1,37**

Таблица: Оценки нашего метода

Предварительное вычисление	Время вычисления, с
нет	9,60
есть	**0,06**

Таблица: Зависимость времени работы от наличия предвычисления

Предварительное вычисление на порядок уменьшает время работы алгоритма.

Заключение

Полученные результаты:

- Предложен метод сравнения текстовых деревьев, учитывающий как структурные, так и семантические их различия.
- Показано, что такой метод сравнения лучше отражает значительные различия в текстовых деревьях, чем используемый до этого.
- Успешно оптимизировано время подсчёта расстояния.

Перспективы исследований:

- Подбор оптимальной языковой модели для моделирования семантического сходства предложений в вершинах деревьев.
- Применение и исследование полученного критерия в задаче иерархической суммаризации.