Знания-Онтологии-Теории (ЗОНТ-2025)

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

Соболевский Федор Александрович[1], Воронцов Константин Вячеславович[1,2]

[1]*Московский физико-технический институт, Керченская улица, 1А, корп. 1, г. Москва, 117303, Россия.*
[2]*Московский государственный университет имени М. В. Ломоносова, Ленинские горы, д. 1, г. Москва, 119991, Россия.*

sobolevskii.fa@phystech.edu, k.v.vorontsov@phystech.edu

Аннотация. *В задачах автоматической иерархической суммаризации текстов возникает необходимость в сравнении текстовых иерархий для оценки качества. Для этого, однако, на данный момент не существует общепринятых метрик, а используемые в предыдущих работах методы оценки качества либо основаны на ручной проверке, либо слабо учитывают структуру и семантику текстовых иерархий. В связи с этим мы предлагаем использовать для сравнения текстовых деревьев расстояние редактирования текстового дерева (text tree edit distance, TTED). Данная метрика основана на расстоянии редактирования дерева, в которой стоимость операций редактирования определяется через семантическое расстояние между текстами в вершинах, аппроксимируемое с помощью выбранной большой языковой модели. TTED призвана отражать прежде всего значимые различия между текстовыми иерархиями, моделируя одновременно как структурные, так и семантические различия между ними. С помощью грамотного подбора модели-энкодера в TTED можно добиться высокого качества сравнения текстовых деревьев между собой. Практическая реализация предложенного нами алгоритма доступна по ссылке на Github:* https://github.com/intsystems/text-tree-distance.

Ключевые слова: текстовые деревья, расстояние редактирования, большие языковые модели, алгоритм Чжана-Шаши.

1 Введение

Текстовые деревья — деревья, в вершинах которых находятся тексты — как структура данных возникают в ряде различных задач: иерархической суммаризации, генерации интеллект карт и др. При разработке методов автоматической генерации подобных структур возникает потребность в оценке качества сгенерированных иерархий. Стандартным подходом к автоматическому оцениванию качества в таких задачах является сравнение с золотым стандартом, созданным вручную экспертом, однако в случае текстовых иерархий для этих целей на данный момент не существует общепринятых метрик. В работах [1–3] для оценки сходства текстовых интеллект-карт применяется функция сходства, основанная на сопоставлении ребер деревьев и сравнении текстов в вершинах при помощи семейства метрик ROUGE [4]. Такой подход, однако, слабо учитывает как конкретику структурных различий между деревьями, так и семантику текстов в них [5]. Это и отсутствие других воспроизводимых метрик для оценки качества генерации текстовых иерархий обуславливает необходимость создания нового метода сравнения текстовых деревьев между собой.

На сегодняшний день существует немало способов сравнивать между собой деревья и тексты по отдельности. Для сравнения структур деревьев применяются такие метрики, как коэффициент

Жаккара [6], расстояние Робинсона-Фоулдса [7] и расстояние редактирования дерева [8]. Для моделирования семантики текстов на сегодняшний день успешно применяются большие языковые модели — в частности, энкодеры на основе архитектуры BERT [9]. Наша работа призвана объединить методы сравнения деревьев и текстов для создания метрики, позволяющей сравнивать текстовые деревья как объекты, объединяющие семантику текстов и структуру деревьев.

2 Теоретические проблемы

Рассмотрим одну из задач, в которой возникает потребность в сравнении между собой текстовых деревьев — иерархическую суммаризацию. Пусть задано множество $\mathcal{S}$ текстов над некоторым словарем. 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 $\mathcal{T}$. Задача иерархической суммаризации — построение отображения $f : D \mapsto T$, строящего иерархическую сводку $T \in \mathcal{T}$ по документу D, минимально отличающуюся от эталонной сводки T^*, построенной экспертом по этой же сводке. Оптимизационная задача в этом случае записывается как

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

Именно здесь и возникает проблема задания адекватной метрики $\rho : \mathcal{T}^2 \to \mathbb{R}^+$ на множестве текстовых деревьев $\mathcal{T}$, отражающей, прежде всего, значимые различия текстовых деревьев. Значимыми мы будем считать различия деревьев по их структуре и по семантике (смысловому содержанию) текстов в их вершинах. В качестве незначительного различия можно выделить, например, перефразирование текстов в вершинах. Нашей целью будет предложить такую метрику, для которой расстояния между деревьями, отличающимися друг от друга только перефразированием, будут в среднем как можно меньше, чем расстояния между деревьями, отличающимися по структуре и/или семантике.

3 Предлагаемый метод.

Предлагается следующая метрика — *расстояние редактирования текстового дерева*, или *TTED* (text tree edit distance). TTED между двумя текстовыми деревьями определяется как расстояние редактирования дерева — наименьшая суммарная стоимость операций редактирования, позволяющих получить из одного дерева другое — со стоимостью операций редактирования, определяемой как семантическое расстояние между текстами в вершинах в случае операции замены вершины и как расстояние от текста в вершине до пустой строки в случае добавления и удаления вершины. Расстояние редактирования с заданными стоимостями операций редактирования можно эффективно вычислять для упорядоченных и неупорядоченных деревьев с помощью алгоритма Чжана-Шаши [8, 10].

Для измерения семантического расстояния между текстами применим большую языковую модель-энкодер (кодировщик) $\mathrm{LM} : S \to \mathbb{R}^n$. С помощью этой модели можно сопоставить текстами некоторые конечномерные векторы (эмбеддинги), расстояние между которыми уже можно измерить с помощью стандартных метрик в $\mathbb{R}^n$. 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 — метрика в $\mathbb{R}^n$. Итого, стоимости операций редактирования в TTED задаются следующим образом:

1. Стоимость замены вершины v на вершину v' — $\rho_n(\mathrm{LM}(s(v)), \mathrm{LM}(s(v')))$;
2. Стоимость удаления/добавления вершины v — $\rho_n(\mathrm{LM}(s(v)), \mathrm{LM}(\lambda))$, где λ — пустая строка.

TTED позволяет, таким образом, сравнивать текстовые деревья, одновременно учитывая их структуру с помощью расстояния редактирования, являющегося основой TTED, и семантику с использованием больших языковых моделей для ее моделирования.

4 Вычислительные эксперименты

Для тестирования TTED с разными моделями-энкодерами в сравнении с базовым методом, использованным в работах [1–3], были измерены расстояния/значения сходства на синтетической выборке, состоящей из деревьев и их модификаций. Средние расстояния между деревьями и их модификациями по перефразированию, структуре и семантике обозначим как $\overline{\rho}_1$, $\overline{\rho}_2$ и $\overline{\rho}_3$ соответственно, средние значения функции сходства — $\overline{\mathrm{Sim}}_1$, $\overline{\mathrm{Sim}}_2$ и $\overline{\mathrm{Sim}}_3$ соответственно. Результаты

тестирования TTED представлены в таблице 1. Для базового метода были получены следующие результаты:

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

Таблица 1: Средние оценки расстояния с помощью TTED с разными моделями-энкодерами

Модель-энкодер в TTED	$\overline{\rho}_1$	$\overline{\rho}_2$	$\overline{\rho}_3$
DistilRoBERTa	3,33	7,76	7,38
SPECTER	1,39	3,70	4,74
MPNet	2,30	7,19	8,06
Дообученная MPNet	1,82	7,71	7,56

По результатам тестирования можно видеть, что структурные различия отражаются базовым методом заметно слабее, чем отличия по перефразированию и семантике. Семантика и перефразирование же отражаются примерно одинаково, что неудивительно, учитывая, что сравнение текстов в вершинах в базовом методе ведётся на уровне лексических единиц с помощью ROUGE. Для сравнения, различия деревьев по структуре и семантике отражаются с помощью TTED заметно сильнее, чем отличия по перефразированию. Более того, для структурных и семантических различий значения расстояний в среднем близки, что позволяет утверждать, что TTED не только более информативно, чем базовый метод, но и сбалансировано отражает различные аспекты различия текстовых деревьев. Все это обосновывает применение метрики TTED для оценки качества в задачах автоматической генерации текстовых иерархий в дальнейшем.

Список литературы

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

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

[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] 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.

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