УДК 004.89, 519.172.1

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

***Ф. С. Соболевский*1*, К. В. Воронцов*1,2**

1Московский физико-технический институт (национальный исследовательский университет)

2Московский государственный университет имени М. В. Ломоносова

В данной работе предлагается и реализуется решение проблемы сравнения текстовых деревьев, то есть деревьев, метками вершин которых являются фрагменты текста. К их числу относятся, например, интеллект-карты на основе предложений (*salient sentence-based mind maps, SSM* [1]) и иерархические сводки документов [2]. Такие структуры данных возникают в задачах автоматической структурированной суммаризации текстовых документов — к примеру, в ряде недавних работ по машинной генерации интеллект-карт по текстам [1], [3-4]. Иерархическая организация текстовой информации в сводке документа является способом улучшить восприятие и запоминание информации текстовой информации [2], [4] и представляется нам потенциальным способом двигаться при освоении новой информации от главного к деталям.

Одной из проблем, возникающих при решении задачи автоматической генерации иерархий из текста, является оценка качества полученных текстовых деревьев. 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. Для сравнения иерархий и текста, однако, на данный момент не существует общепринятых метрик качества, что затрудняет сравнение различных подходов к решению подобных задач. В работах по данной теме текстовые деревья, как правило, сравниваются отдельно по своей структуре как деревья и отдельно с помощью, например, метрики ROGUE [5] как наборы текста [2-3]. Такой подход, однако, не учитывает взаимосвязь между структурой текстового дерева и его содержанием, а применение статистических метрик по типу ROGUE не учитывает семантические сходства/различия текста [6].

Самой распространенной на практике метрикой сходства между деревьями является расстояние редактирования деревьев (*tree edit distance, TED*). При заданной стоимости операций удаления, добавления и обновления вершины расстояние редактирования определяется как минимальная стоимость получения из одного дерева другого с помощью этих трех операций. Такой способ определения расстояния между деревьями стал самым широко применяемым, и для подсчета расстояния редактирования было разработано немало алгоритмов. Один из них — алгоритм Zhang-Shasha, а именно его модификацию для неупорядоченных деревьев [7], мы и применяем для сравнения текстовых деревьев.

В качестве стоимости обновления вершины, то есть замены фрагмента текста в вершине на другой, мы используем степень семантического сходства этих фрагментов, то есть степень их сходства по смыслу. Для автоматической оценки семантической близости мы предлагаем использовать расстояние между эмбеддингами заданных предложений, полученными с помощью заранее выбранной языковой модели. 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 [8-9]. Полученные при помощи нейросетевой модели эмбеддинги можно затем сравнить с помощью некоторой метрики расстояния, например, с помощью косинусного коэффициента (*cosine similarity*) [8]. В таком случае, стоимость обновления метки вершины определяется как Стоимость добавления/удаления вершины естественным образом определяется как степень семантического сходства, вычисленная с помощью той же модель, и пустой строки.

Для усовершенствования полученного алгоритма мы предлагаем несколько модификаций:

* 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 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.
* 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 чем мы определяем коэффициент глубины , позволяющий изменять вес степени сходства между предложениями в вершинах в зависимости от их глубины. Пусть глубина изменяемой вершины , тогда расстояние между предложениями умножается на вес
* 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.

Реализация предложенного нами алгоритма на Python доступна по ссылке на Github: https://github.com/intsystems/text-tree-distance.

**Литература**

1. *Yang W., Honglei G., Jin-Mao W., Zhong S.* Revealing Semantic Structures of Texts: Multi-grained Framework for Automatic Mind-map Generation. // IJCAI. — 2019. — P. 5247–5254.
2. *Christensen J., Soderland S., Bansal G., et al.* Hierarchical summarization: Scaling up multi-document summarization // Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). — 2014. — P. 902–912.
3. *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.
4. *Jain P., Marzoca A., Piccinno F.* Structsum Generation for Faster Text Comprehension // arXiv preprint arXiv:2401.06837. — 2024.
5. *Lin C.-Y.* Rouge: A package for automatic evaluation of summaries // Text summarization branches out. — 2004. — P. 74–81.
6. *Fabbri A. R., Kryscinski W., et al.* Summeval: Re-evaluating summarization evaluation // Transactions of the Association for Computational Linguistics. — 2021. — Vol. 9. — P. 391–409.
7. *Zhang K., Statman R., Shasha D.* On the editing distance between unordered labeled trees. // Information processing letters. 1992 May 25; 42(3): 133-9.
8. *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.
9. *Chandrasekaran, D., Mago, V.* Evolution of semantic similarity — a survey. // Acm Computing Surveys (Csur), 2021, 54(2), pp.1-37.