Text Tree Edit Distance: A Language Model-Based Metric for Text Hierarchies

1st Fedor Sobolevsky
Department of Applied Mathematics and Computer Science
Moscow Institute of Physics and Technology
Moscow, Russia
sobolevskii.fa@phystech.edu

2nd Konstantin Vorontsov
Institute of Artificial Intelligence
M. V. Lomonosov Moscow State University
Moscow, Russia
k.v.vorontsov@phystech.edu

***Abstract*—Text trees as a data structure occur in numerous machine learning tasks like hierarchical summarization and automatic mind map generation. One of the main methods of quality evaluation in these tasks is comparison with reference hierarchies created by experts. The method used so far to compare text hierarchies, as shown in this work, poorly accounts for their structure and text semantics relative to phrasing. To address this issue, we propose a new metric on the set of text trees — text tree edit distance (TTED), based on tree edit distance with semantic distance between texts measured using a large language model. To evaluate how the metric reflects different aspects of text tree difference, we introduce special quality coefficients that reflect the sensitivity of a metric to paraphrasing relative to structural and semantic differences of text trees. Using these coefficients, we conduct extensive testing of the proposed metric and its modifications compared to a baseline used in previous works to compare text hierarchies, which shows that TTED indeed captures significant differences between text trees more accurately than the previously used method. We also provide a practical implementation of TTED for further usage.**

Keywords*—*text trees, mind map, hierarchical summarization, tree edit distance, large language models, Zhang-Shasha algorithm

I. Introduction

In this work, we propose and implement a method for comparing text hierarchies, or text trees, i.e. trees whose vertex labels are text fragments. These include, for example, mind maps (salient sentence-based mind maps, SSM, and key snippet-based mind maps, KSM [1]), hierarchical document summaries [2] and other hierarchies containing textual information. Such data structures occur in automatic structured summarization of text documents [1]–[4]. The hierarchical organization of textual information in a document summary has been proven to be a way of improving information acquisition and retention [2], [3], and we consider it a potential way to efficiently acquire new information from, for example, scientific literature, going from key points to more specific details. A example of a hierarchical summary based on our study is presented in Fig. 1

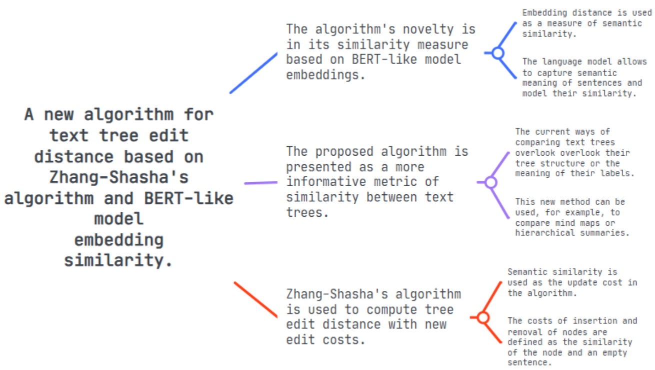
A new algorithm for text tree edit distance based on Zhang-Shasha's algorithm and BERT-like model embedding similarity.
The algorithm's novelty is in its similarity measure based on BERT-like model embeddings.
Embedding distance is used as a measure of semantic similarity.
The language model allows to capture semantic meaning of sentences and model their similarity.
The proposed algorithm is presented as a more informative metric of similarity between text trees.
The current ways of comparing text trees overlook overlook their tree structure or the meaning of their labels.
This new method can be used, for example, to compare mind maps or hierarchical summaries.
Zhang-Shasha's algorithm is used to compute tree edit distance with new edit costs.
Semantic similarity is used as the update cost in the algorithm.
The costs of insertion and removal of nodes are defined as the similarity of the node and an empty sentence.

Fig. 1. An example of a hierarchical summary based on our study

One of the problems in hierarchical text summarization is the evaluation of generated hierarchical summaries. The standard approach to evaluating the quality of a hierarchical summary, as in summarization in general, is to compare the generated summary with a summary of the same document created by an expert. However, for comparing text hierarchies, there currently are no standard metrics, which complicates the comparison of different summarization methods. In existing automatic approaches to evaluating hierarchical summarization, text trees are typically compared separately by structure as trees and as sets of text using, for example, the ROUGE similarity metrics [1], [4]. This approach, however, does not account for the interrelation between structure of text trees and their textual contents, and statistical metrics like ROUGE may not account for text semantics [5]. The low informativity of the similarity function used in [4] will be shown below. This and the absence of other reproducible metrics for hierarchical summarization evaluation necessitate the development of a new metric for this task.

To date, there are many ways to compare hierarchies, which are used in various fields where the task of determining the proximity of trees arises — for example, in computational biology [6], [7]. Among the widely used metrics, one can highlight Robinson-Foulds distance [8], the Jaccard coefficient [9], tree edit distance [10], and others. In this work, we will focus on the latter metric, tree edit distance, as one of the most widely used distance functions for tree comparison [11].

Tree edit distance was first proposed in [10] as the minimal cost of tree editing operations (adding, deleting, and updating a vertex) to obtain one tree from another given the cost of editing operations. The authors, K. Zhang and D. Shasha, also proposed an algorithm for its efficient computation for ordered trees, which now goes by its authors' names. In [12], Zhang and Shasha considered the case of unordered trees; they showed that for unordered trees, the problem of finding the edit distance becomes NP-hard, but for smaller trees, a modification of their algorithm is still applicable. This metric was chosen as the basis for the one proposed in this work due to its interpretability, extensive study, and versatility due to the arbitrariness in choosing the costs of editing operations.

II. Theory

A. Problem Statement

The object of this study is text trees, i.e., trees whose vertices are labeled by text fragments. Let us define this object formally. Let a vocabulary $\mathcal{W}$ and the corresponding set $\mathcal{S}$ of texts over this vocabulary be given:

$$\forall s \in \mathcal{S} \quad s = (w_j)_{j=1}^{|s|}, \quad w_j \in \mathcal{W}. \tag{1}$$

Let us define a text tree $T = (V, E)$, $E \subset V^2$, for each vertex $v \in V$ of which a text $s(v) \in \mathcal{S}$ is assigned. We will denote the considered set of text trees as $\mathcal{T}$. To define an adequate metric on $\mathcal{T}$, let us formulate some requirements for an arbitrary metric on the set of text trees. Let a function of semantic distance between texts be given: $r : \mathcal{S}^2 \to [0, +\infty)$. For vertices $v, v' \in V$ of tree $T = (V, E)$, we will denote $r(v, v') := r(s(v), s(v'))$, and $r(v) := r(s(v), \lambda)$, where λ is an empty string. We will then define a distance function $\rho : \mathcal{T}^2 \to [0, +\infty)$ satisfying the following requirements:

1) $\rho(\cdot, \cdot)$ is a correct metric, i. e. it is symmetrical, positive for unequal arguments and satisfies the triangle inequality.
2) Let $T, T' \in \mathcal{T}$. There exists some non-decreasing function $f : [0, +\infty) \to [0, +\infty)$, such that:
   a) If T' is obtained from T by adding a vertex v to T, then $\rho(T, T') = f(r(v))$;
   b) If T' is obtained from T by deleting a vertex v from T, then $\rho(T, T') = f(r(v))$;
   c) If T' is obtained from T by replacing a vertex v with v', then $\rho(T, T') = f(r(v, v'))$.

B. Metric Informativity

Let us now specify some requirements for the significance of differences between text trees reflected by a distance function. Firstly, it is natural to require that the metric reflect differences between text trees both in *structure* and *semantics* of their text contents. Secondly, an informative metric should weakly react to insignificant differences — for example, to *paraphrasing* of texts in the tree vertices. Therefore, distances between trees obtained from each other by paraphrasing should on average be significantly smaller than distances between trees differing in structure and/or semantics.

Let us formalize these requirements. Let $T \in \mathcal{T}$ be an arbitrary text tree. Let $P(T)$ be the set of trees that can be obtained from T by paraphrasing texts in its vertices, $S(T)$ — the set of trees composed of the same set of vertices with the same texts in them as T (but with different tree structure), $M(T)$ — the set of trees with the same structure as T but with different semantics of the texts in the vertices. For concreteness, let us define the latter set as the set $\mathcal{T}_{\sim T}$ of trees with the same structure as T, excluding the tree T itself and its paraphrases: $M(T) = \mathcal{T}_{\sim T} \setminus (P(T) \cup \{T\})$.

Consider a sample $\mathcal{D}$ of the following form:

$$\mathcal{D} = \{T, T'_1, \dots, T'_p, T''_1, \dots, T''_s, T'''_1, \dots, T'''_m\}, \tag{2}$$

where $T \in \mathcal{T}$, $T'_i \in P(T)$, $T''_j \in S(T)$, $T'''_k \in M(T)$. We introduce the following quality coefficients for metric ρ, defined by the sample $\mathcal{D}$:

$$R_S^{\mathcal{D}}(\rho) = \frac{1}{sp} \sum_{i=1}^{p} \sum_{j=1}^{s} \frac{\rho(T, T'_i)}{\rho(T, T''_j)}, \tag{3}$$

$$R_M^{\mathcal{D}}(\rho) = \frac{1}{mp} \sum_{i=1}^{p} \sum_{k=1}^{m} \frac{\rho(T, T'_i)}{\rho(T, T'''_k)}. \tag{4}$$

Then the task of finding the most informative metric in terms of reflecting significant differences of text trees can be formalized as an optimization problem of the following form:

$$R_S^{\mathcal{D}}(\rho) \longrightarrow \min_{\rho}, \quad R_M^{\mathcal{D}}(\rho) \longrightarrow \min_{\rho}. \tag{5}$$

III. Methods

A. Metric for Text Trees Comparison

Let us analyze the requirements for a metric for comparing text trees formulated in Section II-A. From the requirement that ρ satisfies the triangle inequality, it naturally follows that $\rho(T, T')$ will correspond to the set of tree edit operations of least cost allowing to obtain one tree from the other. It is worth noting that the function ρ defined this way is a correct metric, provided that $f(r(\cdot, \cdot))$ is a metric. The requirements set in II-A are satisfied by tree edit distance with costs of editing operations defined by semantic distance between texts in the vertices with $f(r) = r$. Thus, let us define a new metric on the set of text trees — *text tree edit distance* (TTED). To compute TTED, one can use the Zhang-Shasha algorithm, namely its variants for ordered [10] and unordered [12] trees depending on the specifics of the given application.

To measure semantic distance between texts in tree vertices, we propose to use the distance between vector representations (embeddings) of the texts obtained using some pre-selected language model. Suppose we are given a language model LM : $\mathcal{S} \to \mathbb{R}^n$ that maps text fragments to some finite-dimensional vectors (embeddings). Then we can define for $s, s' \in \mathcal{S}$ the semantic distance between them as $r(s, s') = \rho_n(\mathrm{LM}(s), \mathrm{LM}(s'))$, where ρ_n is a metric in $\mathbb{R}^n$. As a similarity function for normalized embeddings, one can use, for example, cosine similarity S_C, as proposed in [13]. In this case, the corresponding distance function can be defined as $\rho_n(A, B) = \sqrt{1 - S_C(A, B)}$. Also, in this work, the application of standard L_1- and L_2-metrics as ρ_n in TTED will be investigated.

B. Baseline method

For comparison, consider the text tree similarity function used in works [1], [4], [14] to evaluate the similarity of automatically generated mind maps with reference ones. For text trees $T = (V, E)$ and $T' = (V', E')$, the similarity function is defined as:

$$\mathrm{Sim}(T, T') = \max_{P \subset E \times E'} \sum_{(e,e') \in P} (\mathrm{R}(e_0, e'_0) + \mathrm{R}(e_1, e'_1)) . \quad (6)$$

where P is a one-to-one mapping of edges of T to edges of T' and $\mathrm{R}(v, v')$ is the averaged ROUGE-1, ROUGE-2, and ROUGE-L [15] similarity score of $s(v)$ and $s(v')$.

It should be noted that the baseline method uses a similarity function, not a metric on $\mathcal{T}$. Since TTED is a distance function, to compare these methods, one should construct a distance function from $\mathrm{Sim}(\cdot, \cdot)$. This can be done by constructing a pseudometric by analogy with the kernel method [16]:

$$\rho_{\text{base}}(T, T') = \sqrt{k(T, T) + k(T', T') - k(T, T') - k(T', T)}, \quad (7)$$

where $k = \mathrm{Sim}$. This definition of the baseline distance function, firstly, automatically guarantees the following properties:

- Symmetry: $\forall T, T' \in \mathcal{T}\ \rho_{\text{base}}(T, T') = \rho_{\text{base}}(T', T)$;
- Positivity: $\forall T, T' \in \mathcal{T}\ \rho_{\text{base}}(T, T') \geq 0$, and $\rho_{\text{base}}(T, T') = 0 \Leftrightarrow T = T'$.

Secondly, the triangle inequality for ρ_{base} will also hold provided that $\mathrm{Sim}(\cdot, \cdot)$ is a positive definite kernel, but for the similarity function used in [4] this is not the case.

C. Experimental Setup.

To test the proposed method in comparison with the baseline, we compute distances on multiple samples $\mathcal{D}_k$, consisting, according to the definition from Section II-B, of the following elements:

1) a text tree T;
2) trees whose texts in the vertices are *paraphrases* of the corresponding vertices of T — subsample $\mathcal{T}_1$ from $P(T)$ (`paraphrase`);
3) trees that are formed from the vertices of T but with different *structure* — subsample $\mathcal{T}_2$ from $S(T)$ (`structure`);
4) trees that are identical to T in structure but significantly differ in *meaning* of the texts in their vertices — subsample $\mathcal{T}_3$ from $M(T)$ (`meaning`).

The goal is to find among the proposed metrics such a metric ρ for which the quality coefficients $R_S^{\mathcal{D}}(\rho)$ and $R_M^{\mathcal{D}}(\rho)$ will be minimal according to optimization problems (5). Also we introduce the notation $\overline{\rho}_i$ for the average distance between T and trees from subsample $\mathcal{T}_i$. Then a qualitative indicator of the informativity of a metric ρ will be a significantly smaller value of $\overline{\rho}_1$ compared to the values of $\overline{\rho}_2$ and $\overline{\rho}_3$.

D. Experimental Data.

To create a sample for testing the metrics' sensitivity to different aspects of text tree difference, we utilize the generative model DeepSeek V3 [17]. The data generation process consisted of the following stages:

1) Creation of a base tree T using the neural network;
2) Generation of modifications of this tree using the neural network;
3) Manual verification of the generated text trees for compliance with our requirements.

This data generation methodology significantly reduced the time spent on creating the sample and reduced human bias in the data, while maintaining control over their quality. As a result, we created a sample consisting of trees of sizes from 5 to 25 vertices and their modifications of each type discussed above. The test dataset, as well as prompts used to generate it, are provided in the project repository[1].

E. Implementation.

The algorithm for TTED computation was implemented using the Python library `edist`. Language models used to model text semantics were taken from the `sentence-transformers` library[2]. The following language models were used:

- Fine-tuned DistilRoBERTa v2;
- SPECTER [18];
- MPNet [19];
- Fine-tuned MPNet.

The implementation of the baseline method is taken from the official repository of the article [4]. This implementation was adapted to work with the tree format used in our work, but the main logic of the method was left unchanged. All the code allowing to reproduce the results of our experiments is available in our project's repository.

F. Algorithm Modifications

To improve the quality of TTED and compute it efficiently, we implement the following heuristics:

a) Context usage: Often in practice, it is inaccurate to compare texts without considering their context. For example, the sentences "The article discusses it." and "The article discusses the method for text tree comparison." are effectively equivalent if the parent vertex of the first contains the sentence "A new method for text tree comparison is proposed." To address this, in our implementation of TTED we add the option to preliminarily append all sentences from parent vertices as context before the text in a vertex and then compare text embeddings obtained taking this context into account.

b) Precomputation: Repeated computation of embeddings using a neural network model can be very time-consuming for large trees, so in our implementation we precompute embeddings for all texts and distances between them beforehand and then use the precomputed values in the Zhang-Shasha algorithm afterwards.

[1] `https://github.com/intsystems/text-tree-distance`
[2] `https://huggingface.co/sentence-transformers/`

IV. Experimental Results

A. Metric Testing

The results of testing various language models with embedding distance based on cosine similarity in comparison to the baseline method are presented in Table I. The distance estimates obtained using the baseline method and TTED using the MPNet encoder are presented in Fig. 2 and 3 respectively.

TABLE I
Quality coefficients for baseline and TTED with different language models

Model	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Baseline	2.05±0.79	0.96±0.10
TTED with DistilRoBERTa	0.58±0.22	0.53±0.11
TTED with SPECTER	0.69±0.35	0.46±0.14
TTED with MPNet	**0.44±0.12**	0.48±0.11
TTED with fine-tuned MPNet	0.61±0.78	**0.45±0.12**

From Fig. 3 and the values of $R_D^S(\rho)$ in Table I, it can be seen that TTED shows significantly larger distance values for trees differing in semantics and structure than for trees that are paraphrases of each other. This supports the claim that TTED reflects significant differences of text trees much more than insignificant ones. Moreover, for most of the tested encoder models, the distances between trees differing in structure and semantics on average are fairly close. This suggests that TTED balances different aspects of text tree difference. We suggest using the MPNet encoder module in TTED as it produces balanced and low values of $R_S^{\mathcal{D}}(\rho)$ and $R_M^{\mathcal{D}}(\rho)$ as can be seen in Table I.

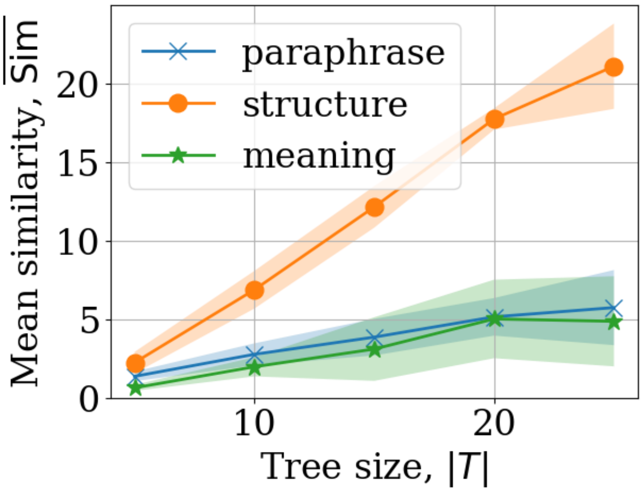

Fig. 2. Baseline similarities for different tree sizes

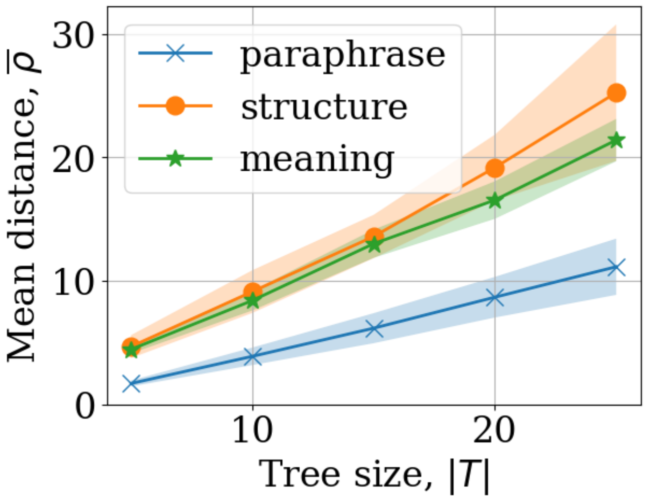

Fig. 3. TTED distances with MPNet encoder for different tree sizes

For comparison, consider the values obtained using the baseline method (Fig. 2, Table I). We see a different picture: first of all, structural differences are reflected much weaker by this method than differences of texts in the vertices, including in wording, hence the value $R_S^{\mathcal{D}}(\rho_{\text{baseline}}) > 1$. This can also be seen in Fig. 2 in much larger values of mean similarity between structurally different trees. Moreover, differences in semantics are on average reflected by the baseline similarity function similarly to differences in phrasing. This can also be seen in the value of $R_M^{\mathcal{D}}$ being close to 1. Here, of course, we should note that $R_S^{\mathcal{D}}$ and $R_M^{\mathcal{D}}$ for the baseline method are calculated using the pseudometric we constructed based on $\overline{\text{Sim}}$, but the values presented in Fig. 2 also allow us to make the same conclusion about low informativity of the baseline similarity function.

B. Testing Algorithm Modifications

Using the TTED version with the MPNet encoder, we experimented with different embedding distances and context usage when computing embeddings. The results of these experiments are presented in Tables II and III. It can be seen that the type of metric used for measuring the distance between embeddings does not visibly affect the value of the $R_M^{\mathcal{D}}$ quality coefficient, but the distance values can differ greatly in order of magnitude depending on the choice of metric. In further experiments, the metric based on cosine similarity will be preferred, as it gives the most interpretable values (with its use, the distance between embeddings is a number from 0 to 1). It should be noted though that it is necessary to normalize embedding vectors in order for this distance function to be a correct metric.

TABLE II
Average TTED values for different embedding distances

$r(x,y)$	$\overline{\rho}_1$, $\lvert T\rvert = 10$	$\overline{\rho}_3$, $\lvert T\rvert = 10$	$R_M^{\mathcal{D}}(\rho)$
$\sqrt{1 - S_C(x,y)}$	3.89±0.71	8.41±0.80	**0.48±0.11**
$\lvert\lvert x-y\rvert\rvert_2$	5.50±1,00	11.89±1.14	**0.48±0.11**
$\lvert\lvert x-y\rvert\rvert_1$	119.70±21.60	259.05±25.12	**0.48±0.11**

From Table III, it can be seen that using context as a heuristic for the TTED computation algorithm indeed gives a slight improvement in the informativity of the metric, which justifies its use in the future. It should be noted though that we cannot guarantee that this modification of TTED satisfies the requirement of being a correct metric on $\mathcal{T}$.

TABLE III

QUALITY COEFFICIENTS WITH AND WITHOUT THE CONTEXT USAGE HEURISTIC

Method	$R_S^{\mathcal{D}}(\rho)$	$R_M^{\mathcal{D}}(\rho)$
Without context	0.44±0.12	0.48±0.11
With context	**0.43±0.19**	**0.35±0.08**

V. CONCLUSION

A. Discussion of Results.

In this work we proposed a new metric for text hierarchy comparison — TTED. This metric has been proven to be more informative in terms of reflecting significant text tree differences than an existing method used to compare text hierarchies. With a quality choice of the encoder model for approximating semantic similarity in TTED, we managed to obtain significantly better values of the proposed quality coefficients than the baseline method; a qualitative investigation of the obtained results also indicates the superiority of TTED. All this justifies the applicability of this metric for comparing text hierarchies; in particular, for evaluating the quality of automatic hierarchical knowledge organization and summarization. The proposed heuristics for TTED can be used to enhance the method.

B. Research Limitations.

Our study of TTED has a number of limitations. First, there are more aspects of similarity of text trees than just structure and semantics. We plan to formalize and investigate such aspects in the future. Also, it should be taken into account that asymptotically the Zhang-Shasha algorithm for unordered trees is quite resource-intensive and could potentially make TTED a non-optimal metric for large text trees; the boundaries of its applicability in terms of the size of the trees compared remain to be determined. We limit our test data to relatively small mind maps on scientific subjects, but the potential scope of usage of TTED is, of course, much broader.

REFERENCES

[1] Y. Wei, H. Guo, J.-M. Wei, and Z. Su, "Revealing semantic structures of texts: Multi-grained framework for automatic mind-map generation.," in *IJCAI*, pp. 5247–5254, 2019.

[2] J. Christensen, S. Soderland, G. Bansal, *et al.*, "Hierarchical summarization: Scaling up multi-document summarization," in *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 902–912, 2014.

[3] P. Jain, A. Marzoca, and F. Piccinno, "Structsum generation for faster text comprehension," *arXiv preprint arXiv:2401.06837*, 2024.

[4] Z. Zhang, M. Hu, Y. Bai, and Z. Zhang, "Coreference graph guidance for mind-map generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 19623–19631, 2024.

[5] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021.

[6] Y. Lin, V. Rajan, and B. M. Moret, "A metric for phylogenetic trees based on matching," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1014–1022, 2011.

[7] F. Pazos and A. Valencia, "Similarity of phylogenetic trees as indicator of protein–protein interaction," *Protein engineering*, vol. 14, no. 9, pp. 609–614, 2001.

[8] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

[9] P. Jaccard, "Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 241–272, 1901.

[10] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM journal on computing*, vol. 18, no. 6, pp. 1245–1262, 1989.

[11] T. Akutsu, "Tree edit distance problems: Algorithms and applications to bioinformatics," *IEICE transactions on information and systems*, vol. 93, no. 2, pp. 208–218, 2010.

[12] K. Zhang, R. Statman, and D. Shasha, "On the editing distance between unordered labeled trees.," *Information Processing Letters*, vol. 42, pp. 133–139, 05 1992.

[13] T. Vrbanec and A. Meštrović, "Comparison study of unsupervised paraphrase detection: Deep learning—the key for semantic similarity detection," *Expert systems*, vol. 40, no. 9, p. e13386, 2023.

[14] M. Hu, H. Guo, S. Zhao, H. Gao, and Z. Su, "Efficient mind-map generation via sequence-to-graph and reinforced graph refinement," *arXiv preprint arXiv:2109.02457*, 2021.

[15] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[16] B. Schölkopf, "The kernel trick for distances," *Advances in neural information processing systems*, vol. 13, 2000.

[17] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[18] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," *arXiv preprint arXiv:2004.07180*, 2020.

[19] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in neural information processing systems*, vol. 33, pp. 16857–16867, 2020.