Threads of Subtlety: Detecting Machine-Generated Texts Through Discourse Motifs

Anonymous ACL submission

Abstract

With the advent of large language models (LLM), the line between human-crafted and machine-generated texts has become increasingly blurred. This paper delves into the inquiry of identifying discernible and unique linguistic properties in texts that were written by humans, particularly uncovering the underlying discourse structures of texts beyond their surface structures. Introducing a novel methodology, we leverage hierarchical parse trees and recursive hypergraphs to unveil distinctive discourse patterns in texts produced by both LLMs and humans. Empirical findings demonstrate that, although both LLMs and humans generate distinct discourse patterns influenced by specific domains, human-written texts exhibit more structural variability, reflecting the nuanced nature of human writing in different domains. Notably, incorporating hierarchical discourse features enhances binary classifiers' overall performance in distinguishing between human-written and machine-generated texts, even on out-of-distribution and paraphrased samples. This underscores the significance of incorporating hierarchical discourse features in the analysis of text patterns. The code and dataset will be available at [TBA].

1 Introduction

The emergence of powerful instruction-tuned large language models (LLMs) (Ouyang et al., 2022; Muennighoff et al., 2023; Kopf et al., 2023) has led to an explosion of machine-generated texts in both offline and online domains. Consequently, discerning the authorship of texts has become a significant challenge, spanning from educational settings to the landscape of online advertising (Extance, 2023; Dalalah and Dalalah, 2023; Gołąb-Andrzejak, 2023). Indeed, many efforts have been made to tackle this issue by constructing corpora of machine-generated and human-authored texts (Dou et al., 2022; Guo et al., 2023; Li et al., 2023) and developing models and benchmarks to tell them apart (Wu et al., 2023; Verma et al., 2023; Su et al., 2023; Chakraborty et al., 2023). The consensus seems to be that while classifiers that make use of the presence of LLM-specific signatures in the generated texts perform relatively well on in-domain texts, their accuracy drops significantly with out-of-domain samples. Furthermore, these detectors can be fooled easily with "paraphrasing attacks" even with in-domain samples (Sadasivan et al., 2023; Krishna et al., 2023).

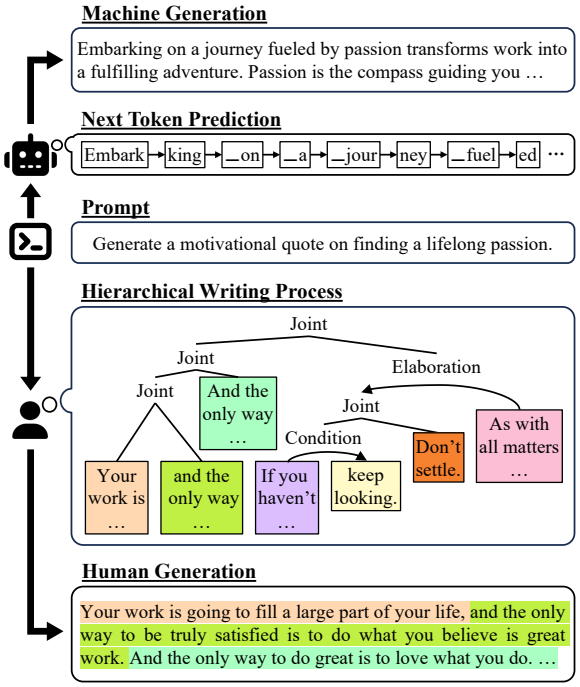


Figure 1: Human writers often employ hierarchical linguistic structures in writing whereas LLMs primarily operate by the sequential next token prediction task.

This raises interesting questions on the underlying nature of human-written texts: "Are there any discernible, unique properties within texts crafted by humans?" and if so, "Might these distinctive signatures manifest at levels beyond surface structure?" Undeniably, how we write varies greatly

depending on the domain and intent addressed by the texts. Consider, for instance, the distinction between academic writing and a casual response to a Reddit post. The former pays much more attention to the logical progression of arguments and overall structural coherence, while the latter is characterized by a more spontaneous and less structured thought process.

On the contrary, LLMs simply generate texts sequentially based on the next token prediction task where the most likely next token in a sequence is predicted given the context of preceding tokens. Figure 1 illustrates this structural difference in the writing processes of LLMs and humans. We note that LLMs may also internally capture these hierarchical structures to some degree as a by-product, especially in their self-attention matrices (Xiao et al., 2021; Huber and Carenini, 2022). However, our interest lies in exploring whether there exist distinctive hierarchical structures that can aid in distinguishing their authorship.[1]

To answer these inquiries systematically, we draw inspiration from discourse analysis in linguistics (Mann and Thompson, 1987), which uncovers structures found within a sentence, between sentences, and among the paragraphs of a document. Specifically, Figure 2 highlights our main approach: Given texts written by humans and LLMs, we construct a hierarchical parse tree for each document. Subsequently, we transform these trees into recursive hypergraphs, allowing us to perform network motif analysis of discourse relationships. The essence of our analysis lies in computing the difference in the distribution of these *"discourse motifs"* between texts generated by machines and those crafted by humans, identifying discernible discourse patterns within each authorship.

We summarize our contributions as follows:

- To the best of our knowledge, we are the first to demonstrate that integrating hierarchical discourse-level features into the authorship detection task results in enhanced and more robust performance.
- Our empirical findings suggest that the robustness against paraphrasing attacks stems from the preservation of higher-level discourse structures, despite significant variations at the sentence level.
- Addressing a critical gap in existing research, where machine-generated texts typically fall within the range of 200 to 500 tokens, we construct a dataset specific to the creative writing domain. This dataset features 20 stories with lengths of up to 8K tokens, providing a more comprehensive perspective on longer-form machine-generated content.

2 Modeling Hierarchical Structures

To explore unique patterns that may extend beyond surface-level structure, it is imperative to employ an expressive framework capable of representing hierarchical structures within texts. Human writers strategically utilize textual structures to systematically convey meaning, thereby augmenting the clarity, coherence, and persuasiveness of their written compositions. In the field of linguistics, the study of document structure is within the domain of discourse frameworks (Mann and Thompson, 1987; Miltsakaki et al., 2004; Lascarides and Asher, 2007). This paper adapts Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) due to its widespread acceptance and the availability of openly accessible pre-trained parsers. In this paper, we make use of the DMRST parser released by Liu et al. (2021).

2.1 Rhetorical Structure Theory

RST assumes that any well-written document can be parsed into a (recursive) discourse tree. In this tree, each leaf node, called an Elementary Discourse Unit (EDU), corresponds to a phrase within a sentence, while higher-level nodes integrate these lower-level representations into more complex structures, such as phrases to sentences to paragraphs. It does so by assigning a discourse label (e.g., elaboration, contrast, cause, effect, etc.) to a relation (i.e., edge) between nodes. The linked nodes can be either "nuclei" (main ideas) or "satellites" (supporting details) depending on their relative importance in the relation. Figure 3 illustrates how a quote from Steve Jobs is segmented into seven EDUs.[2]

2.2 Conversion to Hypergraphs

Within the RST framework, discourse relations emerge not only through the linking of EDUs but also by encapsulating other relations. To naturally represent this process, we opt to utilize recursive hypergraphs, also referred to as "Ubergraphs" in

[1] We provide a more comprehensive comparison with prior studies in Appendix A.

[2] Full RST tree can be found in Figure 8 in Appendix B.

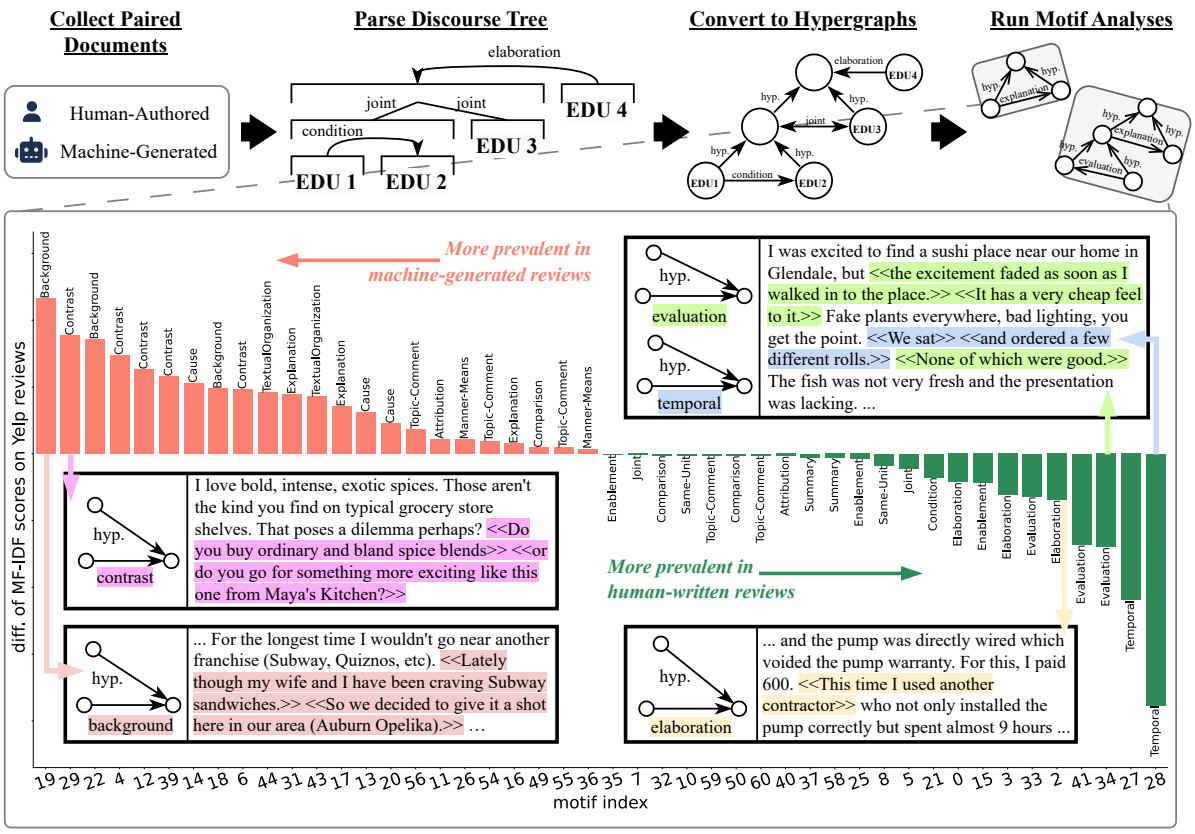


Figure 2: Difference in motif distribution of machine-generated and human-written texts for Yelp domain. Below are the top discourse motifs for each authorship and their corresponding discourse relations with examples. Text enclosed in angle brackets denotes the EDUs involved in the relation.

mathematics (Joslyn and Nowak, 2017), where a discourse relation aligns with a *recursive hyperedge*. This decision is motivated by the potential to leverage various graph analysis techniques through the substitution of the recursive hypergraph with an isomorphic Levi graph.

In short, hypergraphs generalize traditional graphs by allowing edges (or "hyperedges") to connect more than two nodes. The recursive hypergraphs further generalize the standard hypergraphs by allowing hyperedges to contain not only nodes but also other hyperedges. This leads to a hierarchical structure, where hyperedges at different levels of recursion represent relationships at varying levels of abstraction.

To facilitate ease of analysis, the hypergraph can be transformed into a standard traditional graph form by introducing a dummy edge that (depicted as 'hyp.' in Figure 3), signifying the (expanded) hyperedge relation. This enables us to run standard network analysis algorithms as well as train Graph Attention Networks (GATs) (Veličković et al., 2018), leveraging the inherent discourse structure and semantics of documents.

3 Proposed Method: Revealing Key Patterns via *Discourse Motifs*

This paper focuses on finding distinctive structural patterns, namely, hierarchical discourse relations based on the RST framework (§2). Our main investigative approach involves analyzing the distribution of discourse motifs, which are *recurring and statistically significant subgraph patterns* within a larger network. These motifs are often considered as the building blocks of larger networks and can provide insights into the organization and dynamics of the system (Milo et al., 2002; Takes et al., 2018). As illustrated in Figure 2, our proposed approach involves gathering documents written by humans and LLMs, parsing their RST trees, transforming them into hypergraphs, and running motif analyses to understand how individuals (both humans and machines) craft texts. The following subsections illustrate how the discourse motifs are systematically formed (§3.1) and scored using our proposed metrics (§3.2).

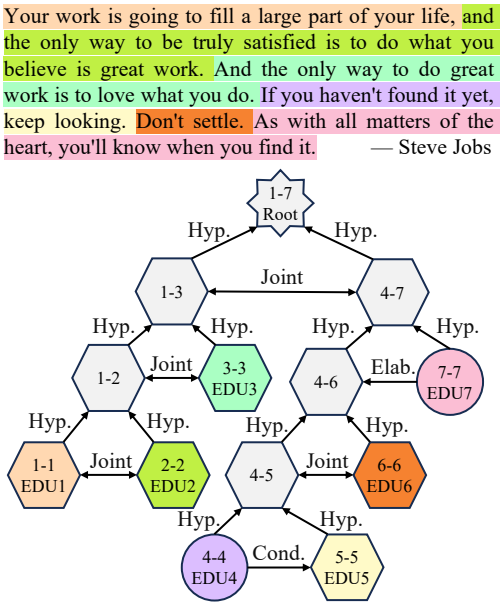


Figure 3: A quote from Steve Jobs and its RST tree (Fig. 8) converted into a hypergraph format. A hexagonal node represents the "nucleus" node, while a circular one denotes the "satellite" node. Each node is labeled with a span of EDU indices that it covers. The 8-point star-shaped node is the root node of the graph, encompassing all subgraphs and EDUs.

3.1 Discourse Motifs as Unions of Triads

One of the standard ways to construct motifs is to generate all possible non-isomorphic subgraphs of a fixed number of vertices. However, as the number of vertices increases, the number of possible patterns increases super-exponentially, rendering their counting extremely inefficient. Complicating matters further, our RST graphs are *directed* with their edges *labeled* with discourse relations, which further increases the complexity.[3]

With this in mind, we start by generating all possible motifs with 3 vertices, i.e., "triads." This is a minimal form of motifs other than the single edge case, "dyad." We select the triads to be the basis motifs as we find that when converting a recursive RST tree into a standard graph, it always forms *the union of subgraphs with 3 vertices*. We highlight the proof in Appendix C.

This process results in 69 non-isomorphic motifs. To create larger motifs, we pick two triads from all possible pairs, join them at all possible nodes, and only keep the non-isomorphic ones, resulting in unions of two triads. We term these configurations as "double-triads." We conduct this process one more round with pairs of a (single-)triad and a double-triad to produce a "triple-triad."[4] We note that depending on how these multiples of triads are joined together, we can form discourse motifs with varying numbers of vertices such as 4, 5, and 7. Table 1 shows the number of identified motifs for each type. Graphical examples of motifs can be found in Appendix L.

Single-Triads	Double-Triads	Triple-Triads
69	592	2,394

Table 1: Number of non-isomorphic discourse motifs.

3.2 Metrics for Analyses

Once we gather potentially useful discourse motifs, we simply count the number of isomorphic subgraphs for each motif across all documents in the datasets. However, as the graph isomorphism test is (possibly) NP-intermediate, the process requires heavy engineering feats to maximize computational efficiency, such as hashing the subgraphs (Shervashidze et al., 2011) and multiprocessing the datasets.[5] Afterward, we can identify distinctive motifs by examining their difference distributions based on their normalized counts (i.e., frequencies) (§3.2.1) as well as calculating our proposed metric termed "motif frequency-inverse document frequency (MF-IDF)" measure (§3.2.2). The rationale behind devising the MF-IDF scoring lies in pinpointing and prioritizing significant motifs, as counting the entire motif distribution is inefficient. An analogy can be drawn to crafting a localized tokenization scheme for LLM pre-training, wherein tokens represent motifs, and texts serve as hypergraphs in our context.

3.2.1 Motif Difference Distributions

As a simple measure, we consider the difference distribution between features of motifs present in graphs. Specifically, we utilize their *motif frequencies*, $\mathcal{F}(\cdot)$ and *weighted average depths*, $\mathcal{D}(\cdot)$ for (document) graphs generated by machines ($\mathbb{D}_{\text{machine}}$) or authored by humans ($\mathbb{D}_{\text{human}}$):

$$\mathcal{F}(\mathbb{D}_{\text{diff}}) = \mathcal{F}(\mathbb{D}_{\text{machine}}) - \mathcal{F}(\mathbb{D}_{\text{human}}) \quad (1)$$

$$\mathcal{D}(\mathbb{D}_{\text{diff}}) = \mathcal{D}(\mathbb{D}_{\text{machine}}) - \mathcal{D}(\mathbb{D}_{\text{human}}) \quad (2)$$

[3] It is worth noting that some discourse relations can be bidirectional as well.

[4] As we are joining motifs at their nodes, the total number of nodes for double-triads and triple-triads will be smaller than 6 and 9, respectively.

[5] The code, as well as the processed datasets for experiments, will be released for ease of reproducibility.

These distributions represent the discrepancies of motifs between the two classes of authorship and can be computed across different domains of texts as well.

Motif Frequency (MF). For each motif m in a set of identified motifs M, we compute motif frequency by counting its occurrences in a graph g and normalize it by the total number of motifs in the graph. The frequency is then averaged over the corresponding dataset $\mathbb{D}$:

$$\mathcal{F}(m, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{g \in \mathbb{D}} \frac{\text{count}(m, g)}{\sum_{m' \in M} \text{count}(m', g)} \quad (3)$$

Weighted Average Depth (WAD). To better capture the hierarchical nature of discourse graphs, we also calculate the average depths of each motif m appearing in a graph g. Let $S(m, g)$ be a collection of subgraphs of a graph g that are isomorphic to a motif m. The depth of a motif m in a graph g is measured by the *mean position*

$$\bar{d}(m, g) = \sum_{g' \in S(m,g)} \sum_{v \in V_{g'}} \frac{d(v, v_{root})}{|S(m, g)||V_{g'}|} \quad (4)$$

where $V_{g'}$ denotes the set of nodes of a subgraph g' and $d(v, v_{root})$ denotes the distance from a node v to the root v_{root}, calculated in g. It is then weighted by the count of the corresponding motif and normalized by the total number of counts:

$$\mathcal{D}(m, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{g \in \mathbb{D}} \frac{\text{count}(m, g) \cdot \bar{d}(m, g)}{\sum_{m' \in M} \text{count}(m', g)} \quad (5)$$

This effectively measures the average position of each discourse motif appearing in the graphs.

3.2.2 Motif Frequency-Inverse Document Frequency (MF-IDF)

Inspired by the term frequency-inverse document frequency (TF-IDF) measure from information retrieval domain (Sparck Jones, 1972), we propose "motif frequency-inverse document frequency" (MF-IDF), noting that a motif (m) can be considered as a vocabulary of documents that are graphs ($g \in \mathbb{G}$) in our case:

$$\text{MF-IDF}(m, g, \mathbb{G}) = \mathcal{F}(m, \{g\}) \cdot \text{IDF}(g, \mathbb{G})$$

where:

$$\mathcal{F}(m, \{g\}) = \frac{\text{count}(m, g)}{\sum_{m' \in M} \text{count}(m', g)}$$

$$\text{IDF}(g, \mathbb{G}) = \log \frac{1 + |\mathbb{G}|}{1 + |g \in \mathbb{G} : m \in g)|}$$

4 Experimental Setup

Following the generation of three sets of discourse motifs for single-, double-, and triple-triads (§3.1), we evaluate them using the MF-IDF metric and retain those with scores surpassing at least one standard deviation. This yields a total of 207 motifs, with a detailed breakdown presented in Table 2.

Single-Triads	Double-Triads	Triple-Triads
31	96	80

Table 2: Number of selected discourse motifs with MF-IDF scores exceeding the threshold of at least one standard deviation.

Section 4.1 outlines the utilization of our proposed features (§3) within both the baseline models and analyses; and Section 4.2 describes the datasets designated and proposed for experimentation.

4.1 Baseline Models

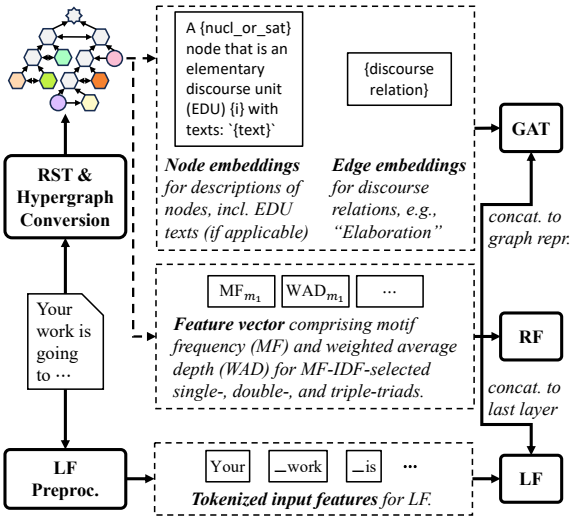


Figure 4: A data flow diagram illustrating how various types of input features are fed into the three baseline models: Graph Attention Network (GAT), Random Forest (RF), and Longformer (LF).

To evaluate the usefulness of discourse motifs, we integrate them into three baseline models for authorship detection (§5.1): Random Forest (RF) (Breiman, 2001), Graph Attention Network (GAT) (Veličković et al., 2018), and Longformer (LF) (Beltagy et al., 2020). Figure 4 depicts an overall process of various types of input features being handled by the different models with varying numbers of model parameters and granularity of inputs. RF is chosen to evaluate the baseline performance solely based on the discourse motif features without considering the semantic content of the texts.

GAT receives the EDU-level representation of the texts as initial node embeddings, along with the discourse edge labels as edge embeddings.[6] The discourse motif features are then appended to the mean-pooled representation of the graph. LF is initialized with pretrained weights[7] and fed with the entire sequence of texts, in addition to the discourse motif features concatenated to its [CLS] token.

We follow Li et al. (2023) in selecting LF and its pretrained weights, but with three distinctions: (i) training a single classifier across all ten domains in the DEEPFAKETEXTDETECT dataset, (ii) using a reduced amount of training data ($\approx 56\%$ of the original set) to address data imbalance, and (iii) refraining from adjusting the decision boundary over ten domain-specific classifiers. Our objective is to create a general classifier, assessing discourse motifs across different model architectures, rather than optimizing for each domain.

Formality Scorer. To identify a relationship between formality style of texts and discourse hierarchies (studied in §5.3), we make use of a publicly available sentence-level formality scorer (Babakov et al., 2023) that is finetuned on a dataset for formality style transfer task, GYAFC (Rao and Tetreault, 2018). The model is based on a ROBERTA-BASE model (Zhang et al., 2020).

4.2 Datasets

Our experiments are conducted on two existing benchmark datasets and one new dataset (TENPAGESTORIES) that we created for more in-depth analysis.

HC3-ENGLISH. Guo et al. (2023) published the first corpus on human writer vs. ChatGPT comparison. It comprises 24K *paired* responses from human authors and ChatGPT for mostly question-answering (QA) style prompts across 5 domains: Reddit-ELI5, medicine, finance, open QA, and Wikipedia-CS.

DEEPFAKETEXTDETECT. Li et al. (2023) proposed DEEPFAKETEXTDETECT dataset which consists of 448K *mostly unpaired* human-written and machine-generated texts from 10 diverse datasets including news article writing, story generation, opinion writing, etc. The machine-generated texts are from 27 mainstream LLMs from 7 sources such as OpenAI and LLaMA. Out of ten domains in the dataset, we present results on the following five considering the diversity and sizes of domains: review writing (YELP), story writing (WP), argument writing (CMV), news summarization (XSUM), and writing descriptions for scientific tables (SCI_GEN).

TENPAGESTORIES (Ours). Existing datasets for the authorship detection task primarily contain short documents of 200 to 500 tokens, limiting the capture of long-term discourse relations; to overcome this limitation, we construct exceptionally lengthy generations based on 20 fictional stories on Project Gutenberg published in early November 2023.[8]

Our generated content is prepared in three distinct settings:

1. **Unconstrained**: We provide the first human-written paragraph of a story and instruct the LLM to iteratively generate continuations of the story.[9]
2. **"Fill-in-the-gap"**: We mask $N \in 1, 3, 5$ paragraphs between a preceding human-written paragraph and a subsequent human-written paragraph.
3. **Constrained "fill-in-the-gap"**: Similar to (2), but the masked paragraph(s) now include the first and last sentences of the corresponding human-written paragraph(s), providing more guided contexts.

These settings are designed to enable the observation of long-term discourse patterns while varying the constraints applied to the original contexts. We generate content up to 8K tokens (≈10 A4 pages) in an iterative manner, continuing from the previously generated texts.[10]

5 Results

We present findings from three sets of experiments:

1. Evaluation of the utility of discourse motifs in the authorship detection task including the "paraphrase attack" scenario (§5.1)
2. Investigation of structural variations after paraphrasing (§5.2)
3. Identifying the relationship between formality and hyperedges (§5.3)

[6] Texts are embedded by using Sentence-BERT (Reimers and Gurevych, 2019) model, ALL-MINILM-L6-V2.

[7] ALLENAI/LONGFORMER-BASE-4096, available on https://huggingface.co.

[8] https://www.gutenberg.org

[9] Due to the limited input length of the existing discourse parser, we did not use the unconstrained generations in experiments.

[10] More detail on the construction can be found in Appx. F.

5.1 Human vs. Machine Authorship Detection

In this section, we report F1 scores for the binary classification task, following the experimental setup detailed in Section 4.1.

HC3 and DeepfakeTextDetect. From Table 3, we can see that incorporating motif information consistently enhances classification performance across various base encoders, underscoring the effectiveness of discourse motifs as auxiliary information for capturing deeper linguistic structures beyond surface lexicons. Specifically, when the GAT model leverages the overall hierarchical discourse structure, notable performance gains are observed (e.g., 0.67 → 0.73 on HC3). Additionally, given the LF model's proficiency in comprehending lengthy texts, the inclusion of discourse motifs offers explicit structural insights, further enhancing the performance. Even on the paraphrased out-of-domain test set ("OOD-Para"), where samples come from "unseen domains" and "unseen models," augmenting the detector with discourse motifs yields significant improvements over the baselines.

Models	HC3	DeepfakeTextDetect		
	Test	Test	OOD	OOD-Para
RF (All Motifs)	0.55	0.57	0.48	0.60
RF (Motifs)	0.55	0.58	0.49	0.61
GAT	0.67	0.68	0.66	0.53
GAT+Motifs	0.73	0.72	0.67	0.54
LF	0.97	0.90	0.74	0.60
LF+Motifs	**0.98**	**0.93**	**0.82**	**0.62**

Table 3: Overall detection F1 scores of models on the benchmark datasets.

TenPageStories. In a more open-ended exploration, we deploy the trained LF model augmented with discourse motifs to analyze our TenPageStories dataset.

Models	FIG			Constrained FIG		
	1	3	5	1	3	5
LF	0.30	0.40	0.45	0.59	0.50	0.55
LF+Motifs	**0.43**	**0.65**	**0.71**	**0.71**	**0.69**	**0.70**

Table 4: Overall detection F1 scores on TenPageStories where "FIG" refers to the "fill-in-the-gap" setting. The numbers 1, 3, and 5 indicate the number of paragraphs to be generated (or "filled in").

First, we check how the different generation settings affect the detection performance (Table 4). As expected the detection model performs better with longer generated texts (i.e., 1, 3, and 5 paragraphs). Also, when the first and last sentences of the human-written texts are additionally provided in the prompt (i.e., "constrained FIG"), the corresponding generations are easier to detect. We note that the models' f1 scores are low as they produce a lot of false negatives, i.e., predicting actual human-written texts to be machine-generated. However, we can observe that the addition of a small discourse vector significantly improves the performance.

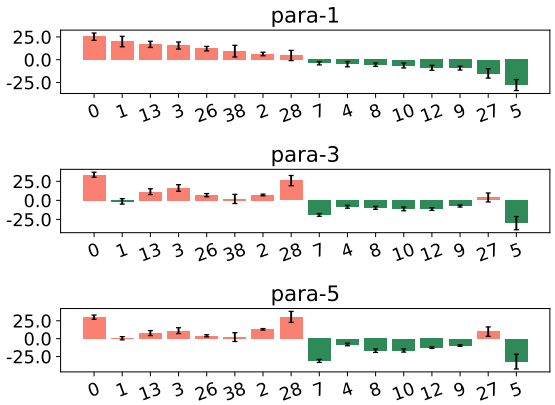


Figure 5: Difference distribution of motifs for TenPageStories under the *fill-in-the-gap* settings. The x-axis represents unique indices of single-triads while the y-axis shows the difference in motif frequency (scaled by 1e-3) of machine-generated and human-written texts for each motif. Discourse motifs indexed at 0 (Elaboration), 5 (Joint), and 28 (Temporal) seem to be useful in distinguishing the two groups.

Similarly, Figure 5 elucidates the distinct motif distributions within TenPageStories across various settings. We can see that the motifs indexed at 0 (Elaboration) and 5 (Joint) seem useful in detecting machine- and human-written texts, respectively. This may imply that human authors tend to construct narratives with a higher prevalence of Joint relations, indicative of more evenly branching structures, in contrast to the patterns found in texts generated by LLMs. Another noteworthy observation is the heightened occurrence of the motif indexed at 28 (Temporal) with increasing text length to generate.

5.2 Structural Variations after Paraphrasing

In this experiment, we look at the discourse structures of original and paraphrased documents regardless of their authorship. We aim to validate our hypothesis that sentence-level paraphrasing induces greater variability in the lower segments of

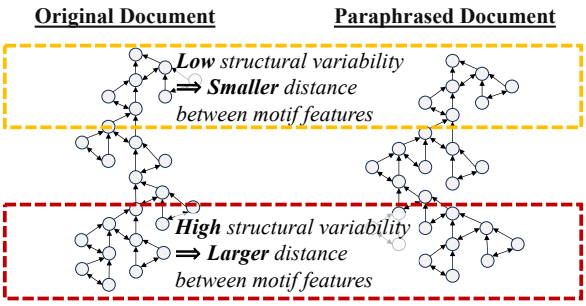


Figure 6: We posit that sentence-level paraphrasing typically leads to greater variability in lower segments of discourse structures compared to higher segments near the root node.

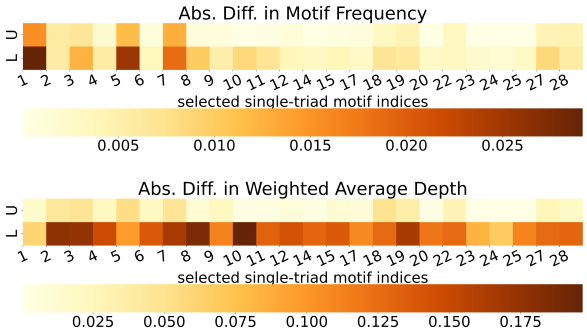


Figure 7: Two heatmaps illustrating the absolute differences of normalized motif counts (above) and weighted average depth (below) between motifs present in discourse graphs of original and paraphrased texts. A graph is further divided into two separate subgraphs (i.e., "upper (U)" and "lower (L)") depending on the distance from the root node.

discourse graphs, as these segments correspond to the individual EDUs and their aggregation into sentences. Given the assumption that the overall higher-level discourse structure remains relatively intact, we anticipate that the upper segments will display less structural variability (c.f. Figure 6). To this end, we calculate the discourse motif features (§3.2.1) for both upper and lower segments of the graphs corresponding to each pair of original and paraphrased documents. Here, the upper segment denotes the neighborhood graph of the root node within an edge distance of 3. Similarly, the lower segment encompasses the union of neighborhood graphs of individual EDU nodes within one edge distance. We average the motif features over the original and paraphrased documents and compute the absolute distance between the two groups.

Figure 7 depicts the results, delineated by motif frequency and weighted average depth. Notably, both features indicate that the lower segments of graphs display a greater absolute distance compared to the upper portions, thereby supporting our hypothesis.

5.3 Formality Scores and Hyperedges

Concerning text styles, we note that the most noticeable disparity among the various text domains appears to lie in the formality dimension. Thus, we computed the formality score for every text generated by humans and machines (c.f. §4.1). Also, for every corresponding document graph, we count the (normalized) frequency of hyperedges for all motifs present in the graph. This is because the hyperedge relations represent higher-level abstractions of text and are thus linked to the discourse hierarchy. From there, we compute the Pearson correlation coefficient between the scores and frequencies. As shown in Table 5, the human-written texts over all domains exhibit weak to moderate correlation with the formality style while machine-generated ones seem to show no correlation. This hints that when humans write more structured texts in terms of discourse, the contents are more formal.

Machine-Generated	Human-Written
0.08	0.39

Table 5: Pearson's R between formality scores and frequency of hyperedges.

5.4 Other Experiments

We conduct preliminary analyses on discourse motifs and investigate the correlation between document graph shapes across different domains in Appendices E and I, respectively.

6 Conclusion

As texts generated by LLMs become increasingly challenging to differentiate from those authored by humans, our approach involves identifying discernible, unique properties inherent in human-crafted texts. We posit that these distinctive signatures manifest at levels beyond mere surface structure and thus opt for looking into their hierarchical discourse structures. Viewing the structures as recursive hypergraphs and conducting motif analyses on them, we find that these motifs are useful in the authorship detection task and can highlight the subtle structural differences in the two author groups depending on the domains of texts and their corresponding nature. Future plans include extending this approach to long documents by merging multiple document graphs and incorporating topological information beyond discourse.

7 Limitations

In our study, the validity of our experiments and subsequent findings relies on the parsed discourse trees generated by an existing parser. It is important to acknowledge the potential for using alternative discourse frameworks like Segmented Discourse Representation Theory (Lascarides and Asher, 2007) though finding or training a robust parser is a challenge. Despite employing a state-of-the-art model for RST parsing, it is crucial to recognize that parsed results may still be suboptimal due to the inherent difficulty of hierarchical discourse parsing. This challenge is exacerbated by the limited scale of existing datasets used for model training and the inherent ambiguities present in discourse relations. In this aspect, an important future direction would be to build a more robust discourse parser using current LLMs.

In our experimental setup, motif analyses were carried out using single-, double-, and triple-triads. It is worth considering that longer and potentially significant patterns may exist beyond these sizes. However, expanding the motif sizes within our current computational approach is not feasible. Nevertheless, there is potential to capture the distributions of longer patterns through approximation techniques, such as subsampling or leveraging deep learning methodologies.

In the experiment investigating the correlation between formality scores and the number of hyperedges (Section 5.3), it is important to acknowledge a limitation: the formality scorer had been trained on short sentence-level inputs and may not be adequately suited for assessing the formality of longer documents spanning multiple paragraphs.

Furthermore, our study highlights the need for more comprehensive analysis within the TENPAGESTORIES dataset. Exploring the correlation between discourse structures and specific linguistic features tailored to creative writing, such as tagged events between named entities, presents an intriguing avenue for future research.

8 Ethical Statement

While our research aims to distinguish linguistic features in human-written and machine-generated texts to improve authorship detection, it is important to acknowledge the potential for these features to be utilized in developing LLMs capable of generating texts that more closely resemble human-authored ones. To address this concern, we advocate for the implementation of transparent reporting practices, ethical review processes, responsible use policies, and collaboration among stakeholders to ensure the ethical development and deployment of LLMs.

Throughout the paper, we have referenced datasets and tools utilized in our experiments, ensuring they originate from open-source domains and do not pose any conflicts with the public release or usage of this paper. Our results are also consistent with the licensing terms of the open-source domain from which the datasets and tools were sourced. We also note that our constructed dataset, TENPAGESTORIES, stems from fictional ebooks available in the public domain and contains no information that names or uniquely identifies individual real people, nor does it include any offensive content.

We acknowledge the use of Grammarly and ChatGPT 3.5 for correcting any less fluent sentences but not for generating new content.

References

Kenza Amara, Rex Ying, Zitao Zhang, Zhihao Han, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. 2022. Graphframex: Towards systematic evaluation of explainability methods for graph neural networks.

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model-generated text: Is ChatGPT that easy to detect? In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 14–27, Paris, France. ATALA.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Daria Beresneva. 2016. Computer-generated text detection using machine learning: A systematic review. In *International Conference on Applications of Natural Language to Data Bases*.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection.

Amrita Bhattacharjee and Huan Liu. 2023. Fighting fire with fire: Can chatgpt detect ai-generated text?

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Megha Chakraborty, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Shreya Gautam, Tanay Kumar, Krish Sharma, Niyar Barman, Chandan Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Counter Turing test (CT2): AI-generated text detection is not as easy as you may think - introducing AI detectability index (ADI). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2206–2239, Singapore. Association for Computational Linguistics.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 148–155, Toulouse, France. Association for Computational Linguistics.

Doraid Dalalah and Osama M.A. Dalalah. 2023. The false positives and false negatives of generative ai detection tools in education and academic research: The case of chatgpt. *The International Journal of Management Education*, 21(2):100822.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12763–12771.

Andy Extance. 2023. Chatgpt has entered the classroom: how llms could transform education. *Nature*, 623(7987):474–477.

Edyta Gołąb-Andrzejak. 2023. The impact of generative ai and chatgpt on creating digital advertising campaigns. *Cybernetics and Systems*, page 1–15.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. Mgtbench: Benchmarking machine-generated text detection.

Patrick Huber and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2394, Seattle, United States. Association for Computational Linguistics.

Cliff A Joslyn and Kathleen E. Nowak. 2017. Ubergraphs: A definition of a recursive hypergraph structure. *ArXiv*, abs/1704.05547.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *ArXiv*, abs/2307.11729.

Andreas Kopf, Yannic Kilcher, Dimitri von Rutte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Rich'ard Nagyfi, ES Shahul, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. *ArXiv*, abs/2304.07327.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Frederick Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023a. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023b. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023c. Check me if you can: Detecting chatgpt-generated academic writing using checkgpt.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164,

Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected?

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. 2011. Weisfeiler-lehman graph kernels. *J. Mach. Learn. Res.*, 12:2539–2561.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2023. Red teaming language model detectors with language models.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412, Singapore. Association for Computational Linguistics.

Frank W. Takes, Walter A. Kosters, Boyd Witte, and Eelke M. Heemskerk. 2018. Multiplex network motifs as building blocks of corporate networks. *Applied Network Science*, 3.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Ting-Hao 'Kenneth' Huang, and Dongwon Lee. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts?

Christoforos Vasilatos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. 2023. Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2023. Ghostbuster: Detecting text ghostwritten by large language models.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2023. A survey on llm-generated text detection: Necessity, methods, and future directions.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Lingyi Yang, Feng Jiang, and Haizhou Li. 2023. Is chatgpt involved in texts? measure the polish ratio to detect chatgpt-generated text.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Neng H. Yu. 2023. Gpt paternity test: Gpt generated text detection with gpt genetic inheritance. *ArXiv*, abs/2305.12519.

Amir Zeldes. 2016. rstweb-a browser-based annotation interface for rhetorical structure theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *ArXiv*, abs/1905.12616.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2022. Distillation-resistant watermarking for model protection in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5044–5055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Comparison with prior work

With the advent of recent LLMs, the need, as well as the challenges in distinguishing machine-generated texts, has risen significantly. Various existing methodologies encompass fine-tuning LLMs (Guo et al., 2023; Liu et al., 2023b,c; Li et al., 2023) with supervised datasets, potentially utilizing techniques like contrastive learning (Liu et al., 2023a; Bhattacharjee et al., 2023), adversarial learning (Shi et al., 2023; Yang et al., 2023), or human-assisted learning (Uchendu et al., 2023; Dugan et al., 2023). Additionally, there are zero-shot methods (Beresneva, 2016; Vasilatos et al., 2023) that leverage statistical features of texts or intermediate values within LLMs; watermarking techniques (Zhao et al., 2022; Kirchenbauer et al., 2023) that introduce "green tokens" or embed "secret keys" as vectors to generated outputs; and prompt-based approaches that make use of (another) LLM as detectors (Zellers et al., 2019; Koike et al., 2023; Yu et al., 2023; Bhattacharjee and Liu, 2023). A comprehensive overview of these approaches can be found in Wu et al. (2023).

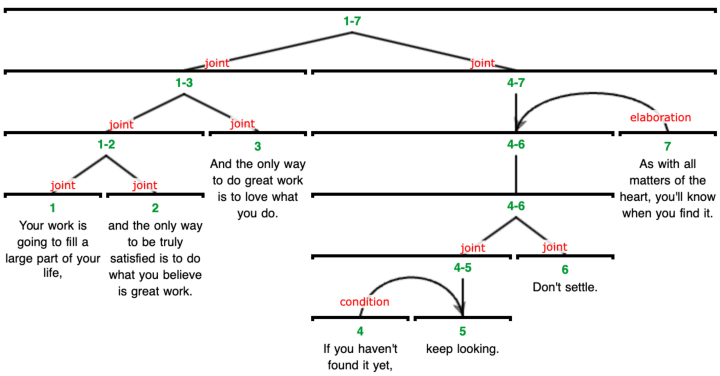
While the relevant recent studies on the detection task primarily adopt surface-level features, such as distributions of n-grams and part-of-speech (POS) tags, we approach the problem from another direction: using hierarchical discourse features within texts. This direction is reminiscent of the work by Corston-Oliver et al. (2001), who explored a comparable path in the pre-LLM era. Their study investigated the efficacy of the branching features observed in syntactic parse trees to determine the origin of texts, particularly whether they were generated by a machine translation model. Our work considers discourse-level features spanning multiple sentences and paragraphs (as opposed to sentence-level POS features).

The existing challenges in the field encompass several facets. First, there are pronounced out-of-distribution (OOD) challenges (Li et al., 2023; Antoun et al., 2023), wherein the detectors struggle when confronted with texts that fall outside the learned distribution. Second, the detectors are prone to potential attacks (Shi et al., 2023; He et al., 2024), such as paraphrasing attacks (Sadasivan et al., 2023; Krishna et al., 2023), where machine-generated texts undergo further paraphrasing to alter the distribution of lexical and syntactic features. This dynamic renders detectors reliant on surface-level features and watermarking technology ineffective. Finally, the inherent ambiguities present in the two groups of texts have become *more nuanced* over time, making it progressively challenging to distinguish them. In this paper, we touch upon these three challenges by showing that the addition of discourse features (i) improves the detection of OOD samples and (ii) trains a more robust classifier against paraphrased attacks. Furthermore, our analyses, utilizing discourse network motifs, shed light on the nuanced distinctions within the hierarchical structures of the two text categories.

B An example of RST tree for the Steve Job's quote

1. Your work is going to fill a large part of your life,
2. and the only way to be truly satisfied is to do what you believe is great work.
3. And the only way to do great work is to love what you do.
4. If you haven't found it yet,
5. keep looking.
6. Don't settle.
7. As with all matters of the heart, you'll know when you find it.

(a) A quote by Steve Jobs, segmented into seven elementary discourse units



(b) A corresponding discourse tree following Rhetorical Structure Theory

Figure 8: Steve Job's motivational quote and its corresponding RST tree.

The RST tree in the figure is drawn with a web tool created by Zeldes (2016).

C Recursive hypergraphs to their standard traditional graph form

In this section, we introduce how recursive hypergraph representations of RST trees can be transformed into standard traditional graph forms. Suppose an RST tree consists of m EDUs $\{\text{EDU}_1, \cdots, \text{EDU}_m\}$ with their relations such as $(\text{EDU}_1, \text{EDU}_2)$ and $((\text{EDU}_1, \text{EDU}_2), \text{EDU}_3)$ (see EDU1, EDU2, EDU3 in Figure 8 for example). Note that in this section we ignore edge labels for better readability. A recursive hypergraph representation of the given RST in Figure 8 is a tuple (V, E) where a set of vertices $V = \{\text{EDU}_1, \cdots, \text{EDU}_7\}$ and a set of edges $E = \{e_{1-2} = (\text{EDU}_1, \text{EDU}_2), e_{1-3} = (e_{1-2}, \text{EDU}_3), \cdots, e_{1-7} = (e_{1-3}, e_{4-7})\}$.[11] To facilitate ease of analysis, the hypergraph can be transformed into a standard traditional graph form as follows.

Definition 1 *A standard traditional graph form (V', E') of a recursive hypergraph representation (V, E) of an RST tree is defined by $V' = V \cup E$ and $E' = \{(u, w) | (u, w) \in E \text{ or } u \in w\}$.*

This graph is thought of as an "extended" Levi graph of a recursive hypergraph (Joslyn and Nowak, 2017).[12] We call it simply the transformed graph of an RST tree if there is no ambiguity.

In the definition of the transformed graph, each relation in an RST tree constitutes a triangle graph. A triangle graph is a graph with three vertices where there is at least one edge between two vertices. For example, a relation $r = (x, y)$ (x and y are either an EDU or an edge) is transformed to the triangle graph $V' = \{r, x, y\}, E' = \{(x, y), (x, r), (y, r)\}$. Now, we generalize this observation as follows.

Theorem 1 *The transformed graph of an RST tree only consists of the union of triangle graphs.*

Proof sketch by the mathematical induction. Before we prove, assume that every finite RST can be represented by a huge, single relation in a recursive manner. For example, the RST in Figure 8 can be represented by

$$(((\text{EDU}_1, \text{EDU}_2), \text{EDU}_3), (\text{EDU}_7, ((\text{EDU}_4, \text{EDU}_5), \text{EDU}_6))). \tag{6}$$

Consider an RST tree $(\text{EDU}_1, \text{EDU}_2)$ as an initial case. Then, the transformed graph is a triangle graph.

Suppose x that is an RST tree with $n - 1$ EDUs is the union of triangle graphs. Then, an RST tree (x, EDU_n) or (EDU_n, x) is also a graph with the union of triangle graphs by the definition.

More generally, suppose that an RST tree with less than n EDUs forms the union of triangle graphs. For an RST tree z containing n EDUs, this RST can be divided into an RST tree x with $n - k$ EDUs and an RST tree y with k EDUs, i.e., $z = (x, y)$. Since both transformed graphs of x and y are triangle unions by the assumption and a relation (x, y) can be transformed into the triangle graph as discussed above, the RST z can be also transformed into a new graph with triangle unions.

[11] We can consider bi-direction by adding a reverse order of an edge.

[12] Different from the definition of Levi graphs in (Joslyn and Nowak, 2017), we allow cycles in the extended Levi graphs.

D Calculation of motif features

Figure 9 illustrates the calculation of the two features, motif frequency and weighted average depth, of a discourse motif with index 18. This motif is one of the selected single-triads that exhibit MF-IDF scores (§3.2.2) surpassing at least one standard deviation.

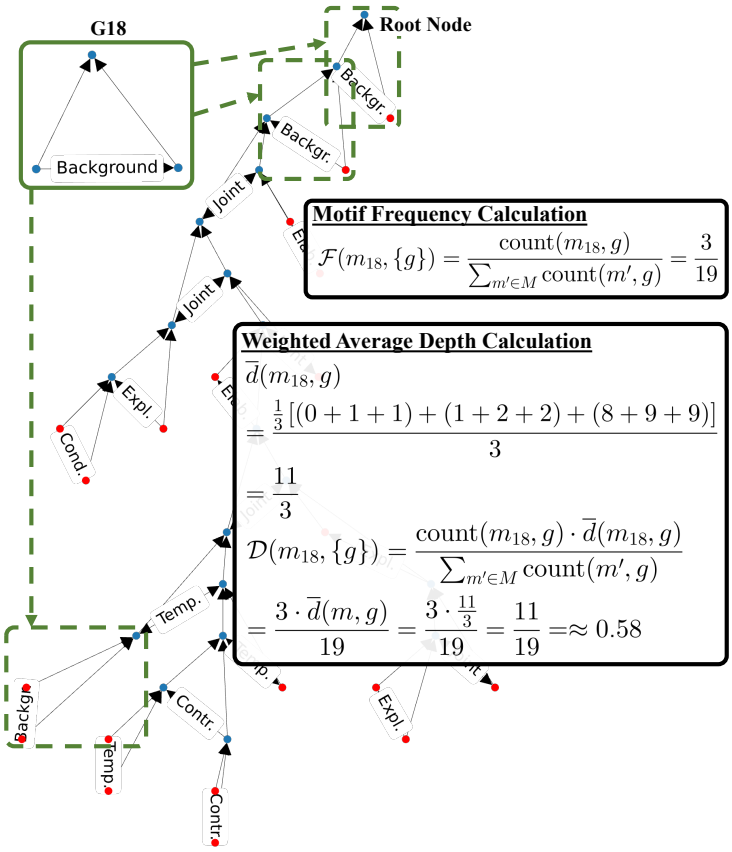


Figure 9: An example of calculating motif frequency and weighted average depth for the single-triad m_{18}.

E Preliminary analyses on discourse motifs

We begin our preliminary analyses by examining the relationship between the proposed discourse motifs and influential edges identified by the GAT explainer (§E.1). Subsequently, we calculate the difference distributions of motifs across five domains to emphasize their distinctive features (§E.2).

E.1 Correlation between Motifs and Influential Edges for Explanation

As a way to evaluate the appropriateness of using discourse motifs, we explore how they are related to influential edges in GATs computed by an explanation method. Specifically, we utilize explainability techniques such as "GNNExplainer" (Ying et al., 2019) and "GraphFramEx" (Amara et al., 2022) for graph classification, aiming to assess their correlation with our proposed motif features. These methods elucidate the predictions of GNNs by identifying the nodes or edges whose masking has the most significant impact on the predicted outcome. Since our graphs consist of directed and labeled edges, we require an explanation method tailored to edge-based features for GAT. Thus, we employ "AttentionExplainer" from GraphFramEx, which computes edge masks using the attention coefficients generated by GATs trained on the binary classification task.

Note that the influential edges refer to edges with high masking values[13] generated by a GAT explainer as described in (§4). We then calculate a correlation between frequencies of selected discourse motifs based on MF-IDF scores and those of influential edges. To this end, we define two random variables as follows. Let X be the number of single-triads and let Y be the number of influential edges for each graph. Samples

[13]The most influential edges whose masking values are larger than 0.99 are considered in $[0, 1]$.

of X and Y are collected from each graph in the datasets (HC3-ENGLISH, DEEPFAKETEXTDETECT), and the Pearson correlation coefficient r_{XY} is calculated.

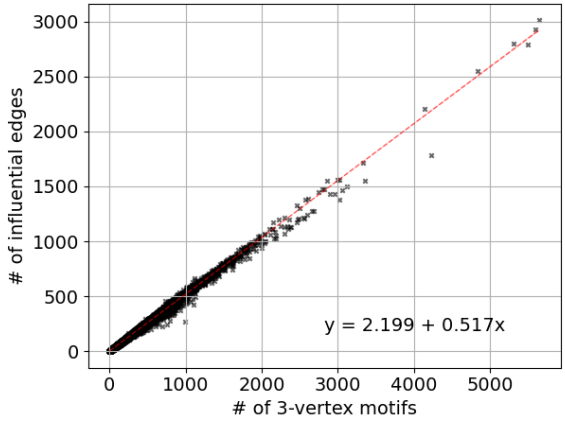


Figure 10: Correlation between selected single-triads with high MF-IDF scores and influential edges computed by a GAT explainer on DEEPFAKETEXTDETECT dataset.

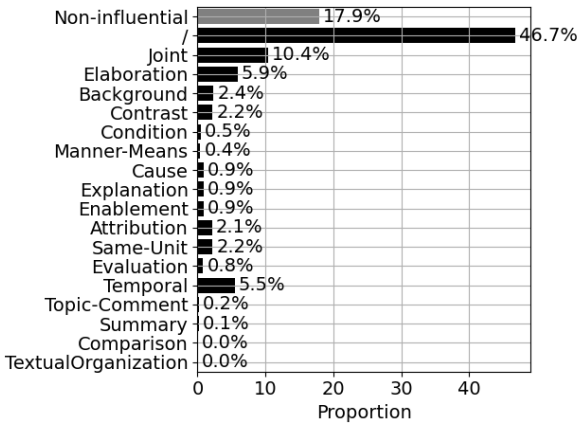


Figure 11: Proportions of labels of influential edges. Note that "Non-influential" refers to edges whose mask values are under the threshold. Note also that the value "0.0%" in this figure does not indicate the zero, but a very small positive number like "1e-5%."

Figure 10 shows that frequencies of selected single-triads and influential edges are highly correlated with $r_{XY} \approx 0.99$, indicating that the discourse motifs may be useful features in determining the authorship of texts. Figure 11 shows the proportion of edge labels (i.e., discourse relations) of influential edges. We note that the label "/" in the figure denotes the hyperedge relation ("hyp.").

It is noteworthy that while the motifs are motivated by the discourse framework rooted in linguistics, the notion of influential edges stems from a machine learning-driven, data-centric approach. The strong correlation observed between the two is particularly interesting.

E.2 Difference Distributions of Motifs across Domains

Figure 12 shows bar plots for the difference distribution of single-triad motifs across the five domains of texts. We note that large-scaled plots can be found in Figure 16.

The mean and standard deviation of differences ($\mathcal{M}(\mathcal{D}_{\text{diff}})$) are is 0.00012 and 0.0015, respectively. Notably, machine-generated texts tend to feature more motifs with Background relations, while human-written texts exhibit an inclination towards Temporal discourse relations in domains such as review writing (YELP) and story generation (WP), and contain more Joint relations in domains that demand formality or deeper logical depths (i.e., argument writing (CMV) and news summarization (XSUM). We note that Joint discourse relation signifies cases where the corresponding child trees are of equal importance (i.e., "nuclei") and thus branched evenly. The texts in table description generation (SCI_GEN) are interesting in

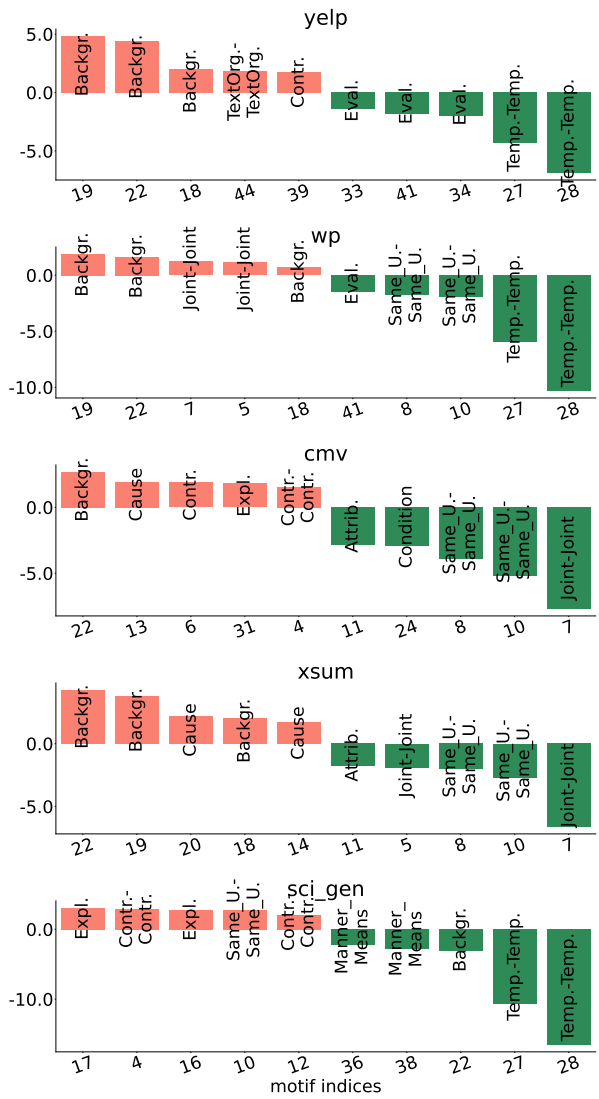


Figure 12: Difference in single-triad motif distribution of machine-generated and human-written texts for five different domains in DEEPFAKETEXTDETECT dataset. Y-axis represent the difference in MF-IDF scores scaled by 1e-3. Full-scaled versions including a larger number of motifs and plots for double- and triple-triads can be found in Appendix K.

that the human-written descriptions (taken mostly from academic papers in NLP and machine learning) mainly highlight the changes in performance and this is captured as Temporal, Manner, or Means relations. Table 6 shows some grounded examples of texts for prevalent motifs in each domain.

F Constructing "TenPageStories" dataset

The "TenPageStories" dataset consists of stories with approximate lengths of 8,000 tokens or 10 A4 pages. It was constructed using an iterative method of calling the OpenAI's GPT-4-1106-PREVIEW model. In total, we conducted 552 request calls, resulting in a total expenditure of 55 USD.

The construction process works by continuously adding the text generated by the model in one call to the prompt of the next call. The iterative method was used for all three generation settings: (1) unconstrained generation, (2) "fill-in-the-gap", and (3) constrained "fill-in-the-gap". Pseudo-code and examples for the three generation settings are provided below. Note that for the constrained fill-in-the-gap example, three different completion calls are made for an iteration: the first completes paragraph 1, the second generates paragraph $2...n-1$, the third completes the last paragraph n, where n is the number of paragraphs to be masked per iteration. The exception to this process is when $n = 1$, then one completion call is made to

complete the paragraph based on its first and last sentences.

Algorithm 1 Unconstrained Generation

```
function UNCONSTRAINEDGENERATION(promptInstruction)
    prompt ← promptInstruction
    while length(prompt) < 8000 tokens do
        generatedText ← generate text using LLM based on the prompt
        prompt ← prompt + generatedText
    end while
    return prompt
end function
```

Algorithm 2 "Fill-in-the-gap"

```
function FILLINTHEGAP(promptInstruction, numMasked)
    prompt ← promptInstruction + firstParagraph
    for each group of numMasked+1 paragraphs from the second paragraph onwards do
        generatedText ← generate text using the LLM to fill the first numMasked paragraphs in the group
        prompt ← prompt + generatedText + lastParagraphInTheGroup
    end for
    return prompt
end function
```

Algorithm 3 Constrained "fill-in-the-gap"

```
function CONSTRAINEDFILL(promptInstruction, numMasked)
    prompt ← promptInstruction
    for each group of numMasked paragraphs do
        firstParagraph ← generate text using the LLM to fill the first paragraph in the group based on its first sentence
        prompt ← prompt + firstParagraph
        middleParagraphs ← generate text using the LLM to fill all paragraphs in the group besides the first and last
        prompt ← prompt + middleParagraphs
        lastParagraph ← generate text using the LLM to fill the last paragraph based on its last sentence
        prompt ← prompt + lastParagraph
    end for
    return prompt
end function
```

1. **Free Generation Example Prompt**:

> Take a look at the story and generate text according to the instructions in brackets []. Only return the generated parts. Here is the story:
> *Text from previous generation*
> *Text from previous generation*
> *Text from previous generation*
> [Generate text here]

2. **"Fill-in-the-gap" Example Prompt (Fill 1 paragraphs)**:

> Take a look at the story and generate text according to the instructions in brackets []. Only return the generated parts. Here is the story:
> *Original paragraph text*
> *Text from previous generation*
> *Original paragraph text*
> [Generate paragraph of approximately 94 tokens here]
> *Original paragraph text*

3. **Constrained "Fill-in-the-gap" Example Prompt (Fill 3 paragraphs)**:

> Take a look at the story and generate text according to the instructions in brackets []. Only return the generated parts. Here is the story:
> *Text from previous generation*
> *Text from previous generation*
> *Text from previous generation*
> *Original first sentence of paragraph 1*
> [Complete preceding paragraph using approximately 67 tokens]
> [Generate paragraph of approximately 129 tokens here]
> [Complete following paragraph using approximately 87 tokens]
> *Original last sentence of paragraph 3*

G Details on experimental setup

Our experiments were conducted on a workstation equipped with 4 NVIDIA RTX A5500 GPUs, an AMD Ryzen Threadripper PRO 5975WX 32-core CPU, and 1TB of RAM. When running experiments, we primarily adhered to the default hyperparameter configurations of the baseline models and model optimizer, with precise details available in the accompanying code repository: `[TBA]`.

We report that RF, GAT, and LF models underwent training for 1 epochs (≈ 10 minutes), 10 epochs (≈ 7.5 hours), and 4 epochs (≈ 14.5 hours), respectively, with the best-performing checkpoint chosen based on validation dataset performance.

We also report that parsing RST structures and extracting discourse features across the entire dataset required approximately 9.5 hours, utilizing the multiprocessing capabilities of CPU cores to their fullest extent. We note that these features only need to be computed once offline.

H More examples of grounded motifs in texts

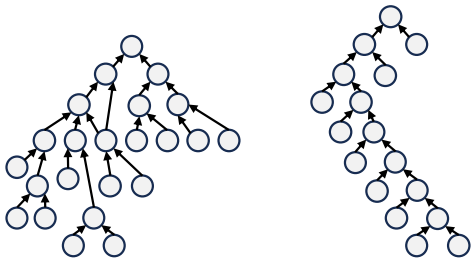
Table 6 shows more examples of texts featuring discourse motifs with high MF-IDF scores.

Domains	Machine-Generated	Human-Written
Review Writing (YELP)	[Motif #19 (Background)] I recommend reservations. [[We were there at 8:30pm, food was excellent with very friendly staff]] <<that made us feel welcome right away.>> The wine list had something …	[Motif #28 (Temporal)] When Lagasse's Stadium first opened we were excited by the idea of watching the big game [[while at the same time enjoying some of those great dishes]] <<we watched Emeril make on TV ->> perfect combination, right? Well, as it turns out …
	[Motif #22 (Background)] Buffets began popping up along the Strip late last decade <<when a number of casinos decided>> [[they wanted to boost traffic by offering some form of cheap dining experience without offending guests with lower cost options like pizza slices or pasta salads.]] While most buffets offer …	[Motif #34 (Evaluation)] [[The stir fry bowl was very fresh, the one thing I noticed was that they probably don't clean their stir fry grill that often between bowls as some of my beef tasted like fish (not good for those with food allergies), but I would assume you could ask them to clean it first.]] <<Could have been a bit more beef in the bowl for the price.>> Not sure if I'd go again.
Story Generation (WP)	[Motif #19 (Background)] 10-year-old Lily eagerly awaited the arrival of hers. She had always been a good girl, always sharing and playing nicely with others. As she sat in her room, [[she wondered]] <<what kind of animal would be assigned to her.>> Would it be a furry puppy? A fluffy kitten? Or something more exotic …	[Motif #28 (Temporal)] I back to the side of the stage and the stage lights come up. <<Suddenly I can see the audience illuminated in parts of the backdrop.>> [[Pretty much everybody - a hundred or so, that is - got quiet after I started arguing with Jon about payment. Now there's clapping, screaming, and I think some lady took off her shirt way back there.]] So apparently the new guy is popular. …
	[Motif #7 (Joint-Joint)] … Elizabeth went on to describe her life, her hopes, and her dreams. <<She wrote about her family, her friends, and her home.>> [[She even included a few sketches of her hometown, which had long since faded away.]] As I read, I felt a strange connection to Elizabeth. Though she lived in …	[Motif #10 (Same_Unit-Same_Unit)] It's not about literal wealth but instead "personal wealth." <<There's a man in the middle of nowhere>> [[living with deer]] that uses up a lot of our Happiness supply. There's a janitor that spends his life cleaning up after others who gets a fresh supply every day.
Argument Writing (CMV)	[Motif #13 (Cause)] For example if you want to know something about your relative or friend but they don't answer your text messages <<because he doesn't use his cellphone>> <<(because there would no longer be any need)>> then what else can you expect him to do except send an email? Well, unfortunately …	[Motif #7 (Joint-Joint)] <<It's already been happening with factory workers, online shopping, etc.,>> [[and with self-driving cars on the horizon and computing technology getting better and better, the long-term job security for many professions is shaky at best.]] This surge in unemployment may even cause a great economic crisis.
	[Motif #6 (Contrast)] To start off, yes, I am Mormon <<and no I am not offended by The Book of Mormon Musical.>> <<In fact, I know a ton of very devout Mormons>> who I saw The Book of Mormon musical in Chicago a couple of weeks ago and it was incredible!	[Motif #24 (Condition)] … but what I'm saying is that they shouldn't be impeached or forced to resign, or face criticism on their ability to lead or shape policy. <<If they are a good policy maker fine.>> <<If they are making the world a better place>> [[and fighting for things that other politicians are afraid to]] then fine.
News Summarization (XSUM)	[Motif #22 (Background)] Hamilton went on to win and now trails Rosberg, who was given a five second penalty in Austria <<following an investigation into what happened>> [[after he returned ahead at Turn One but made contact with his Mercedes teammate as they headed towards Turns Three and Four respectively.]] Rosberg apologized immediately afterwards …	[Motif #7 (Joint-Joint)] … the country's biggest artificial lake. The sub-sea cable will connect to the UK network at Blyth in Northumberland. <<Crucially, the UK will be able to call on the power at short notice.>> [[The energy will be used to manage the growing levels of intermittent wind power on the network.]] It will also be a two-way link.
	[Motif #20 (Cause)] The television personality said she was proud to be Welsh but recognized that many people felt differently <<because there are different views across the United Kingdom>> [[about whether their vote could lead to independence for Wales from Britain or Scotland leaving Britain along with Northern Ireland into an independent EU bloc.]] She noted …	[Motif #11 (Attribution)] … getting rid of fees would help more young people into university, including the disadvantaged? Universities are worried <<that such a switch to direct funding, dependent on government finances,>> [[would put a limit on places and a brake on expansion.]] One of the quiet revolutions of recent years has been the complete removal of limits on student numbers …
Table Descriptions (SCI_GEN)	[Motif #17 (Explanation)] For mobile robots navigating on sidewalks, it is essential to be able to safely cross street intersections. [[Most existing approaches rely on the recognition of the traffic light signal]] <<to determine whether a robot can proceed or stop mid-crossing.>> However in our settings …	[Motif #28 (Temporal)] In this context, all the polarimetric information seems irretrievably mixed. <<A direct model, derived from a simple but original extension of the widespread "multiple scattering model" leads to a high dimensional linear inverse problem.>> [[It is solved by a fast dedicated imaging algorithm that performs to determine at a time three huge 3-D scatterer maps which correspond to HH, VV and HV polarizations at emission and reception.]] It is applied successfully to …
	[Motif #4 (Contrast-Contrast)] <<Recent research on Automatic Chord Extraction (ACE) has focused on the improvement of models based on machine learning.>> <<However, most models still fail to take into account the prior knowledge.>> Unlike the current approaches, prior knowledge is largely ignored in the automatic chord extraction (ACE) problem, …	[Motif #38 (Manner_Means)] The accumulation of litter and plastic debris at the seafloor and the bottom of rivers are extremely harmful for the aquatic life. We propose a solution for monitoring this problem <<using a team of Autonomous Underwater Vehicles (AUVs)>> [[to exchange the recorded video in order to reconstruct the map of regions of interest.]] However, underwater video transmission is a challenge …

Table 6: Grounded examples of texts for some prevalent motifs. Exact shapes of the motifs can be found in Fig. 18.

I Additional experiment: branched vs. chain-like structures

In this experiment, we differentiate texts by domain and assume that the formality of texts is consistent within the domain. We can then look at the varying levels of graph structures per domain. As illustrated in the Introduction (§1), we assume that the linguistic structure of discourse can change depending on factors such as formality, spontaneity, and depth of reasoning in the texts.



(a) Branched document structure

(b) Chain-like document structure

Figure 13: Illustration of document graph structures with different discourse hierarchies. For simplicity, the horizontal edge linking two child nodes is omitted.

One possible dimension of characterizing this difference in structures is by checking whether the structures are evenly branched or follow a more sequential pattern as illustrated in Figure 13.

To quantify this notion, we calculate the average shortest path length (ASPL) of a document graph. For linear or chain-like graphs, the ASPL tends to be relatively short, as nodes are connected in a linear fashion. In contrast, more spread-out graphs will likely have a longer ASPL due to increased distances between nodes. Another closely related metric is the diameter of a graph which is the longest shortest path between any two nodes in the graph. However, as human-written and machine-generated texts are not necessarily paired and could vary a lot in length, we opt for ASPL.

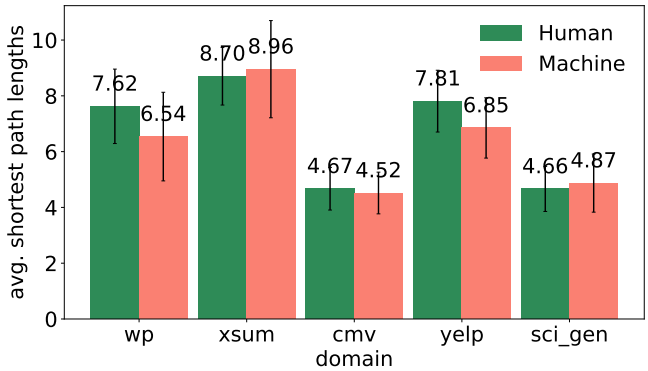
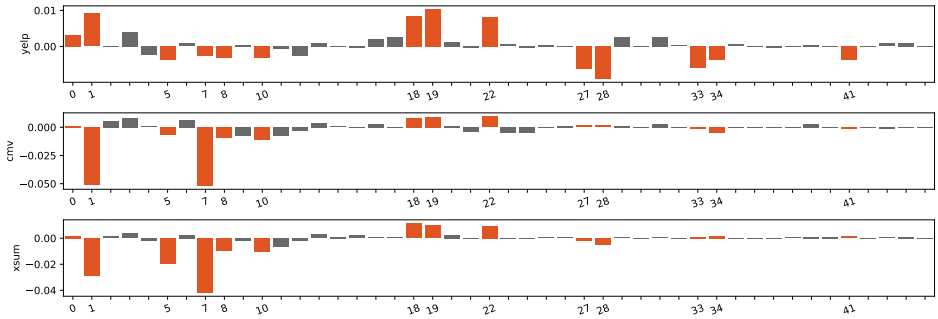


Figure 14: Average shortest path lengths per domain for document graphs. We

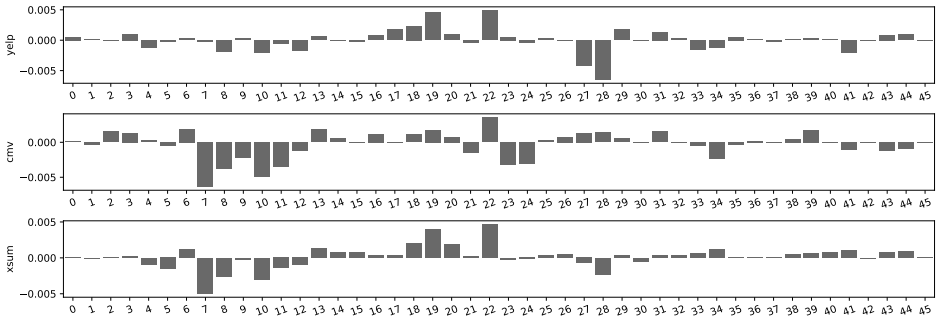
We observe that the news summarization task (XSUM) yields texts with the longest ASPL across both groups. Human-generated texts exhibit longer ASPL in the WP, CMV, and YELP domains. Notably, within these three domains, Joint discourse relations emerge as the most prevalent relation on the human-authored side, affirming our hypothesis regarding the "equal branching" characteristic of this relation (c.f., Fig. 16).

J Difference distributions computed by motif frequency vs. MF-IDF scores

Figure 15 presents two bar plots, illustrating (a) motif frequency and (b) MF-IDF scores of motifs across three domains. Notably, due to the IDF scaling, the latter plot exhibits a slightly less skewed pattern compared to the former plot.



(a) Difference distribution of motifs computed by using their motif frequencies.

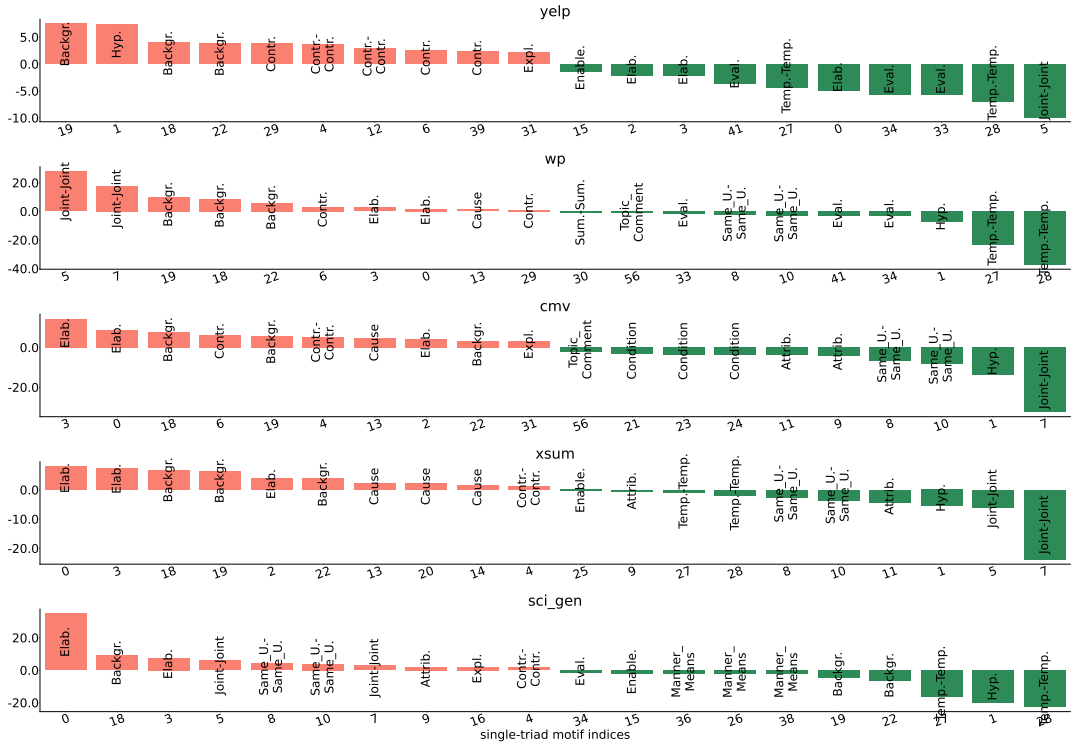


(b) Difference distribution of motifs computed by using their MF-IDF scores.

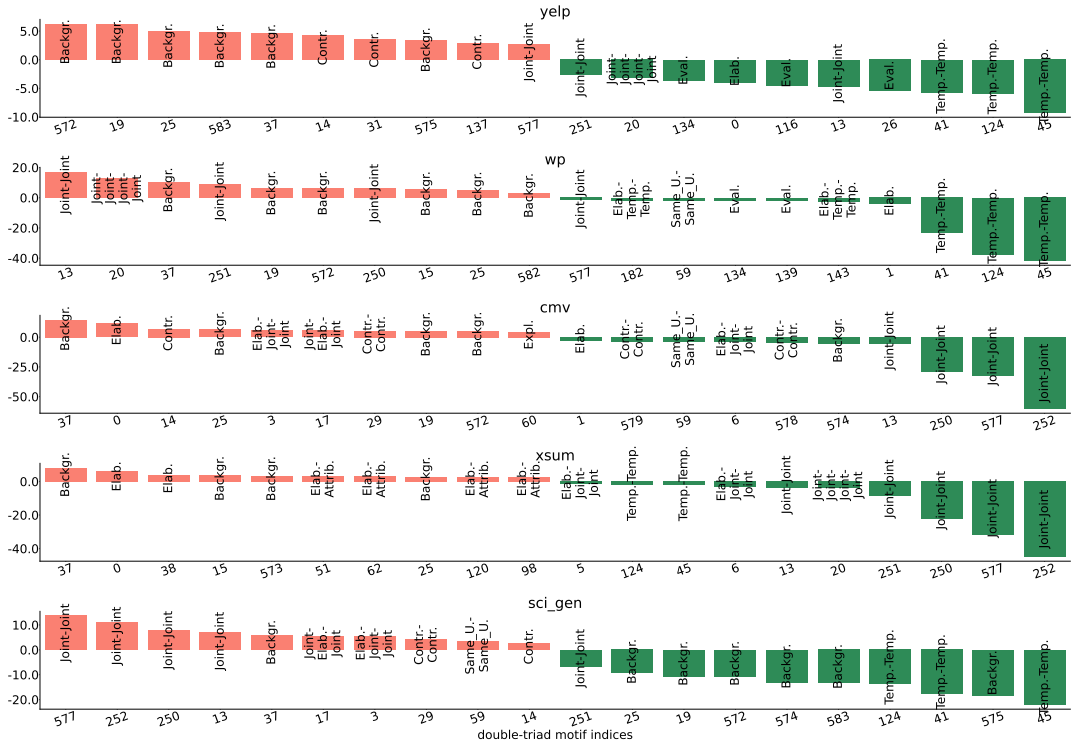
Figure 15: Two ways of computing the difference distributions.

K Large-scale difference distributions

Figure 16 displays two bar plots for single- and double-motifs, depicting MF-IDF difference distributions. Triple-triads are excluded due to label clutter.



(a) Single-triads.



(b) Double-triads.

Figure 16: MF-IDF difference distributions of single- and double-triad motifs. To improve readability, the hyperedge relations have been excluded.

L Graphical examples of identified motifs

Figure 17 illustrates examples of the different types of triads. Similarly, Figure 18 shows all 69 single-triad motifs.

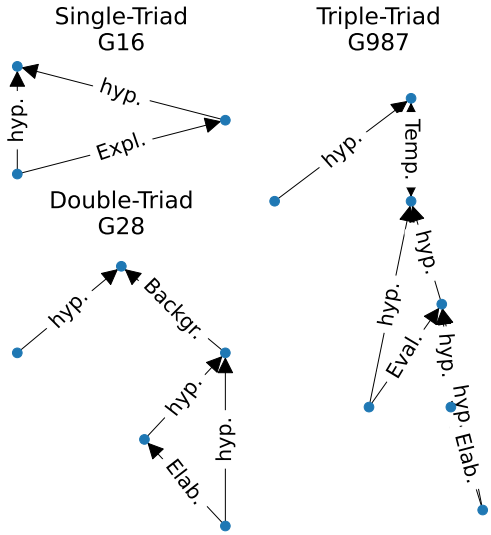


Figure 17: Examples of three types of motifs. More examples of single-triads are drawn in Figure 18.

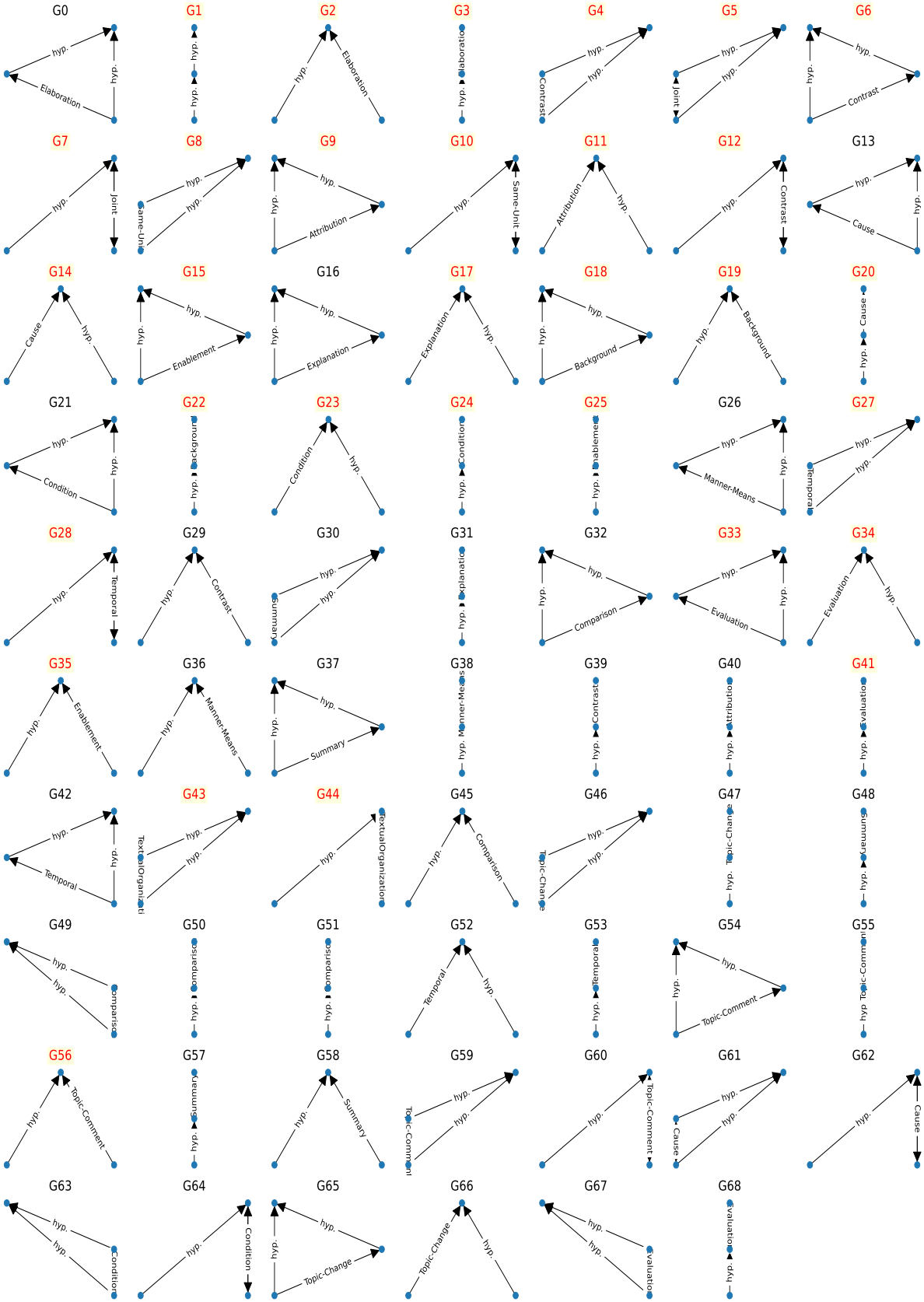


Figure 18: All 69 motifs of size three. The 31 marked motifs are the selected ones that exhibit MF-IDF scores surpassing at least one standard deviation.