Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.7717

Narrative Consolidation: Formulating a New Task for Unifying Multi-Perspective Accounts

Roger Antonio Finger [**UNISINOS** | ***rfinger@edu.unisinos.br***]
Eduardo Gabriel Cortes [**UNISINOS** | ***egcortes@edu.unisinos.br***]
Sandro José Rigo [**UNISINOS** | ***rigo@unisinos.br***]
Gabriel de Oliveira Ramos [**UNISINOS** | ***gdoramos@unisinos.br***]

Graduate Program in Applied Computing, Universidade do Vale do Rio dos Sinos, Av. Unisinos, 950, Cristo Rei, São Leopoldo, RS, 93022-750, Brazil.

Received: 03 March 2026 • **Accepted:** 12 August 2026 • **Published:** 28 August 2026

Abstract Processing overlapping narrative documents, such as legal testimonies or historical accounts, often aims not for compression but for a unified, coherent, and chronologically sound text. Standard Multi-Document Summarization (MDS), with its focus on conciseness, fails to preserve narrative flow. This paper formally defines this challenge as a new NLP task, *Narrative Consolidation*, whose central objectives are chronological integrity, completeness, and the fusion of complementary details, and establishes the resources needed to study it: a formal task definition, an evaluation paradigm that includes a selection-level metric, the *Gospel Consolidation Language Resource* — a benchmark built from the four Biblical Gospels, with 169 canonical events, cross-document alignments, and a manually created reference consolidation — and a suite of reference systems, ranging from a timeline-agnostic centrality method to timeline-aware heuristics and the *Temporal Alignment Event Graph (TAEG)*, a multi-relational graph that explicitly models chronology and event alignment. Benchmarking these systems yields three findings that characterize the task. First, the explicit temporal backbone is the dominant factor: every system granted the canonical timeline raises ROUGE-L F1 from 0.206 to at least 0.81, whereas content-selection sophistication accounts for a far smaller margin. Second, on a fusion-style reference, a simple length heuristic — selecting the longest available account of each event — is remarkably strong (0.947 ROUGE-L F1; 0.648 selection accuracy), outperforming the graph-based selection (0.846; 0.489), which is nevertheless clearly above the random floor. Third, ablations show that the discriminative signal resides in the temporal edges, while intra-cluster lexical similarity is uninformative — a negative result that constrains the design of future models. Together, these results establish Narrative Consolidation as a task distinct from summarization, provide the first reference points for it, and pose an explicit open challenge: to surpass that heuristic with a principled selection mechanism, and to move beyond version selection towards the fusion of complementary details.

Keywords: Narrative Consolidation, Task Formulation, Benchmark Resource, Reference Baselines, Multi-Document Summarization, Temporal Coherence, Temporal Alignment Event Graph

1 Introduction

Multi-Document Summarization (MDS) is typically defined as the task of producing a single, concise summary from a cluster of related documents [Ma *et al*., 2023]. This focus on conciseness, however, is ill-suited for scenarios involving long, complex, multi-perspective narratives, such as witness testimonies in a criminal investigation or overlapping Biblical narratives like the Gospels. In these scenarios, the main objective is not to create a shorter text, but rather to produce a single, *consolidated and unified narrative*. This goal echoes early work in multi-document summarization on "information fusion" [Barzilay *et al*., 1999], but shifts the priority from redundancy removal to narrative completeness. Such a narrative must prioritize three core objectives: strict chronological integrity, completeness in its coverage of events, and the coherent fusion of complementary details from each source. This challenge has been recognized since antiquity with early Gospel harmonization efforts like Tatian's Diatessaron, as documented by the fourth-century historian Eusebius of Caesarea [Eusebius, 1999]. Consequently, the final consolidated text may even be longer than some of the individual source documents, as its value lies in its completeness and coherence, a direct contrast to the brevity-focused goal of standard summarization.

Classic graph-based algorithms focused on conventional summarization like LexRank [Erkan and Radev, 2004] exemplify this limitation. By modeling sentences as nodes and their semantic similarity as edges, these methods inherently ignore chronological flow. This issue is not confined to older methods. Even recent neural and graph-based approaches [Yasunaga *et al*., 2017], while more sophisticated, often share a similar limitation when applied to narrative consolidation. Whether based on Graph Convolutional Networks or large-scale pre-trained transformers like PEGASUS and PRIMERA [Zhang *et al*., 2020a; Xiao *et al*., 2022], their primary objective remains identifying semantic salience for compression, not enforcing a strict chronological narrative. This focus on *what is important* over *what happened when* makes them inherently unsuited for creating a single, coherent, and temporally faithful narrative. These limitations underscore the need to define a new task focused on chronological unification

rather than semantic compression.

This paper argues that unifying multi-perspective narratives constitutes a problem distinct from summarization. It is a task and resource paper: its purpose is to define *Narrative Consolidation* as a new Natural Language Processing (NLP) task, to provide the resources required to study it — a benchmark, an evaluation paradigm, and a set of reference systems — and to report what benchmarking those systems reveals about the nature of the task. Among the reference systems we include the *Temporal Alignment Event Graph (TAEG)*, a multi-relational graph structure that encodes chronology and event alignment explicitly; it serves here to test whether an explicit relational prior helps, not as a method advocated over alternatives.

Our main contributions can be summarized as follows:

- We formalize *Narrative Consolidation*, a new NLP task focused on unifying multiple overlapping narrative accounts into a single, chronologically coherent, and comprehensive text.
- We introduce the *Gospel Consolidation Language Resource*, a new benchmark dataset for this task, including source texts, temporal alignments, and a manually created reference consolidation.
- We propose an evaluation paradigm for the task, including a *selection-level* metric that measures version choice directly and remains discriminative where corpus-level metrics saturate.
- We provide a graded suite of reference systems — a timeline-agnostic centrality method, four timeline-aware heuristics, and the TAEG — together with ablations and a timeline-degradation analysis, establishing the first set of comparison points for the task.
- We report the empirical findings that emerge from this benchmarking, including two results that constrain future work: the dominance of the temporal backbone over content-selection sophistication, and the strength of a simple length heuristic on fusion-style references.

We note that the present study deliberately addresses the *extractive* setting of the task under a known canonical timeline; abstractive approaches to Narrative Consolidation, built upon the same language resource, are investigated in a companion study by the authors [Finger *et al*., 2026].

2 Related Work

Our work builds upon and extends several lines of research in multi-document text analysis, graph-based summarization, and temporal information processing.

2.1 Multi-Document and Timeline Summarization

MDS has evolved significantly from early extractive approaches to sophisticated neural models. Classical methods focused on identifying and fusing similar information across sources [Barzilay *et al*., 1999], while neural approaches like PEGASUS [Zhang *et al*., 2020a] and PRIMERA [Xiao *et al*., 2022] have achieved state-of-the-art results in abstractive summarization. However, these approaches primarily target conciseness rather than comprehensive consolidation. We also note that ordering the selected information is a classical concern within MDS itself, with a dedicated line of research on sentence ordering for multi-document summaries [Bollegala *et al*., 2010].

Timeline Summarization, or Multi-document Temporal Summarization, evolved to incorporate temporal dimensions explicitly. The field has developed two dominant paradigms: date-wise approaches that first identify salient dates and then generate summaries for each date [Martschat and Markert, 2017], and event-based approaches, where the system first clusters sentences into events and then orders these events to form a timeline, not a narrative.

As introduced in Yu *et al*. [2021], Multi-Timeline Summarization (MTLS) generates multiple timelines to address ambiguous or multi-faceted stories. While innovative, MTLS focuses on discovering different storylines within a collection, whereas Narrative Consolidation unifies different versions of the same story.

2.2 Graph-Based Summarization Methods

Graph-based methods have proven highly effective for summarization tasks. LexRank [Erkan and Radev, 2004], which models sentences as nodes and uses centrality measures to identify important content, remains a robust baseline for extractive summarization. More recent work has explored neural graph-based approaches [Yasunaga *et al*., 2017], combining the structural advantages of graphs with the representational power of neural networks.

Our TAEG framework differs fundamentally from traditional graph-based summarization in that the graph structure represents the narrative's temporal and versional relationships rather than semantic similarity between sentences. We apply a centrality algorithm (LexRank) locally at the event level, combining the structural integrity of the TAEG with a proven selection method.

2.3 Temporal Graphs and Dynamic Graph Neural Networks

As highlighted by Santana *et al*. [2023], a central challenge in narrative extraction is linking components through temporal reasoning to form a coherent timeline. While static knowledge graphs can represent relational facts, they fall short in capturing the evolving nature of a story.

This has led to a growing interest in *Temporal Graphs* (or dynamic graphs), which explicitly incorporate time. These graphs are not fixed structures but are often represented as sequences of timed events or interactions, making them a natural fit for modeling narratives. Concurrently, Graph Neural Networks have emerged as a powerful paradigm for learning on graph-structured data. A key research challenge has been adapting these models to operate on dynamic graphs, leading to the development of specialized architectures like Temporal Graph Networks [Rossi *et al*., 2020]. Many of these models are designed to *learn* temporal dependencies from the data itself, often using memory modules to track the evolution of the graph's topology and features [Jiao *et al*., 2025].

Our TAEG framework takes a different philosophical approach. Instead of learning the temporal structure, it leverages a known, canonical timeline as a strong architectural prior. This positions the TAEG not as a model for discovering temporal patterns, but as a scaffold that enforces chronological integrity. This allows other components—like a centrality algorithm or a future GNN—to focus exclusively on content selection and representation within a fixed temporal backbone.

2.4 Biblical and Historical Text Processing

The challenge of creating a single, harmonized narrative from the four Gospels is not new, dating back centuries to theological and scholarly efforts like Tatian's Diatessaron in the second century, which attempted to weave the four accounts into one [Petersen, 1994]. While these manual harmonies aimed to create a unified scripture through conflation, our work approaches the task from a computational linguistics perspective, focusing on extracting and ordering the most representative version of each event to produce a consolidated text.

3 The Narrative Consolidation Task

To establish a rigorous foundation for future research, we formalize Narrative Consolidation as a distinct task, outlining its components, objectives, and evaluation paradigm.

3.1 Formal Definition

The Narrative Consolidation task is defined as follows:

- **Input:** A set of N documents $D = \{d_1, d_2, ..., d_N\}$, where each document d_i is a narrative account of related events. An optional, but often available, input is a canonical timeline T, which is an ordered list of M canonical events $\{e_1, e_2, ..., e_M\}$.
- **Output:** A single document, $D_{consolidated}$, which is a narrative that synthesizes the events from D. The structure of $D_{consolidated}$ must be governed by the canonical timeline T, ensuring that the sequence of presented events is correct. Furthermore, the text must be coherent, fusing information from multiple sources without redundancy, and comprehensive, covering all events for which a description exists in D.
- **Objectives:** The ideal output must satisfy the following criteria:
    1. **Chronological Integrity:** The sequence of events in $D_{consolidated}$ must strictly follow the canonical timeline T.
    2. **Completeness:** $D_{consolidated}$ should cover all canonical events $\{e_1, \ldots, e_M\}$ for which a description exists in at least one document in D.
    3. **Representativeness:** For each event e_j, the version presented in $D_{consolidated}$ should be the most informative and representative among all available versions in D, or, even better, an enriched version with all descriptions combined.
    4. **Redundancy Elimination:** Information common to multiple source documents should appear only once in $D_{consolidated}$.

We remark that treating the canonical timeline T as an input is a deliberate scoping decision: the present study isolates the version-selection subproblem under a known chronology — a setting that genuinely occurs in practice, such as Gospel harmonization, legal case files with an established chronology of facts, or clinical patient histories — so that the contribution of the consolidation mechanism can be assessed independently of timeline induction. Automatically inducing T from the documents themselves is a distinct research problem, further discussed in Section 9.

3.2 Distinction from Related Tasks

Narrative Consolidation is distinct from other multi-document processing tasks, as summarized in Table 1. While it shares roots with Multi-Document Summarization (MDS), its objectives diverge significantly from both standard MDS and its temporally-aware variants, Timeline Summarization (TLS) and Multi-Document Temporal Summarization (MDTS).

	MDS	TLS/MDTS	NC
Goal	Compression	Chrono.Summ.	Unification
Format	Summary	List/Summary	Narrative
Length	Shorter	Shorter	Can be Longer
Chrono.	Secondary	Key Principle	Key

Table 1. Comparison of Narrative Consolidation (NC) with Multi-Document Summarization (MDS) and temporally-aware summarization (TLS/MDTS).

- **Multi-Document Summarization (MDS):** The primary goal of MDS is compression [Ma *et al.*, 2023]. The output is a concise text that captures salient information, but often at the cost of chronological flow. We acknowledge that ordering the summary content is a classical concern in MDS, addressed by a dedicated line of research on sentence ordering [Bollegala *et al.*, 2010]; in MDS, however, chronological ordering is one of several means to improve the readability of a compressed summary, whereas in NC chronological integrity is a constitutive requirement of the task itself. In contrast to the compression goal, NC prioritizes completeness, and the resulting text may be longer than individual source documents to cohesively incorporate all relevant details.
- **Timeline and Temporal Summarization (TLS/MDTS):** These tasks, such as those described by Martschat and Markert [2017] and Mamidala and Sanampudi [2021], incorporate time as a central organizing principle. However, their fundamental goal remains summarization—to produce a concise, chronological overview of events, typically in the form of a list or a series of dated summaries. The output is a *chronology*, not a single, continuous story. NC differs critically in its objective and output: the goal is not a compressed overview but a comprehensive unification, and the output is not a list of events but a single, fluid *narrative*.

3.3 Contradictions, Evolving Facts, and Divergent Perspectives

Realistic multi-document scenarios exhibit not only redundancy across sources, but also *contradictory information* and *divergent perspectives*. Contradictions may arise from the temporal evolution of facts — in breaking news about a plane crash, for instance, successive reports present different casualty counts and candidate causes as the facts "evolve" — from outdated information coexisting with newer data, or from plainly incorrect reports. Divergent perspectives arise when sources frame the same event under opposing viewpoints, as is common in political news.

These phenomena interact directly with the objectives of Narrative Consolidation. In the TAEG representation (Section 4.2), competing accounts of an event concentrate precisely inside the corresponding `SAME_EVENT` cluster, where all versions of that event coexist. A consolidation mechanism for such data would therefore need to go beyond representativeness-based version selection, incorporating, for example, contradiction detection between versions, version timestamping (to prefer the most recent factual state), and stance- or perspective-aware selection or fusion — possibly presenting conflicting versions side by side when no resolution is warranted. The Representativeness and Redundancy Elimination objectives of Section 3.1 would then need to be balanced against *factual faithfulness* to each source.

The Gospel corpus used in this study deliberately controls for these factors: the four accounts are complementary rather than adversarial, with no significant factual contradictions within the adopted chronological framework. We argue that this is a desirable property for a *foundational* benchmark, since it isolates the chronological-unification problem from the orthogonal problems of contradiction resolution and perspective handling. It does, however, limit the direct generalization of our empirical findings to noisier domains such as journalistic or legal corpora; we return to this limitation, and to the construction of a news-domain benchmark exhibiting these properties, in Section 9.

3.4 Evaluation Paradigm

The unique nature of Narrative Consolidation requires a tailored evaluation paradigm. We propose that evaluation directly address the task's objectives along two complementary dimensions:

- **Primary Metric (Temporal Coherence):** A rank correlation measure, such as **Kendall's Tau (τ)**, should be the primary metric to directly evaluate chronological integrity, as it specifically quantifies the agreement between two ordered sequences. In this case, the generated narrative and the canonical timeline.
- **Content Quality Metrics:** Complementarily, the informational fidelity of the consolidated narrative with respect to a reference consolidation should be assessed by content similarity metrics, at both the lexical and the semantic levels. The operational definitions of all metrics employed in this study are consolidated in Section 6.1.

Proposing a specific evaluation paradigm is itself a key step, enabling rigorous and meaningful benchmarking of future models. However, we acknowledge that the development of bespoke evaluation metrics tailored specifically for Narrative Consolidation is a complex challenge that constitutes a significant research direction in its own right. As such, this endeavor falls beyond the scope of the present study. Our approach, therefore, is to utilize the best-suited metrics from the existing literature, recognizing them as effective proxies even if they are not perfectly calibrated for this new task. We posit that this pragmatic evaluation is sufficient for this foundational work, while highlighting the development of more holistic, task-specific metrics as a critical avenue for future research.

4 Reference Systems for Narrative Consolidation

A new task requires reference points against which future work can be measured. This section describes the suite of reference systems we provide with the benchmark. They are organized as a graded ladder along the two factors that plausibly drive performance on the task: access to a canonical timeline, and the sophistication of the per-event version-selection criterion. At the bottom of the ladder sits a timeline-agnostic centrality method, representing how standard extractive summarization approaches the problem (Section 4.1). Above it sit four timeline-aware heuristics, which receive exactly the same canonical timeline but select versions by trivial criteria (Section 4.3). Finally, the Temporal Alignment Event Graph (Section 4.2) combines the timeline with an explicit relational structure over event versions, allowing us to test whether such a structure buys anything beyond the timeline itself. None of these systems is advanced as the solution to the task; their role is to delimit what is easy, what is hard, and what remains open.

4.1 Timeline-Agnostic Reference: A Semantic Similarity Graph

The baseline system represents the standard approach for graph-based extractive summarization, treating the four Gospels as a single multi-document collection [Erkan and Radev, 2004].

4.1.1 Graph Construction

- **Nodes:** Each sentence from all four documents becomes a node in the graph.
- **Edges:** Edge weights are computed as the cosine similarity between the TF-IDF vector representations of any two sentences across the entire corpus.

This results in a dense, undirected, weighted graph where edge weights represent semantic relatedness. LexRank is then applied to this graph to rank and select the top-scoring sentences for the summary.

4.1.2 Inherent Limitation for Narratives

This semantic-only approach inherently ignores the chronological flow of events. An analysis of the baseline's actual output reveals significant temporal disorder. For example, a sentence describing the preparations for the Triumphal Entry appears after sentences detailing Jesus's arrest and even after a post-resurrection account (`https://github.com/neemias8/TAEG`). The following excerpts from the baseline's output illustrate this chronological incoherence:

Excerpt from Baseline Output (Actual Results):

1. Then Jesus said to the chief priests... "Am I leading a rebellion, that you have come with swords and clubs?" (Arrest)
2. When the chief priests had met with the elders and devised a plan, they gave the soldiers a large sum of money... (Post-Crucifixion)
3. As they approached Jerusalem and came to Bethphage on the Mount of Olives, Jesus sent two disciples... (Triumphal Entry)

This example, taken directly from our experimental results, provides empirical evidence that optimizing for semantic centrality alone is insufficient and ill-suited for the Narrative Consolidation task, as it fails to preserve the most fundamental element of a story: its sequence.

4.2 A Structured Reference System: The Temporal Alignment Event Graph (TAEG)

The third reference system constructs a graph by leveraging external knowledge from a chronology file based on [Aschmann, 2022], which identifies 169 canonical events in the Holy Week. Its purpose in this study is diagnostic: by encoding both the chronological order and the version alignment as explicit relations, it lets us ask whether a structured prior over event versions improves selection beyond what the timeline alone provides.

4.2.1 TAEG Construction

The graph is built as a multi-relational structure where nodes represent event versions, not individual sentences.

- **Nodes:** For each of the 169 canonical events, a distinct node is created for each Gospel that describes it. The number of nodes per event thus varies. For instance:
  - The event *The Triumphal Entry*, described in all four Gospels, generates four nodes in the graph.
  - The event *The Institution of the Lord's Supper*, described in the three Synoptic Gospels (Matthew, Mark, and Luke), generates three nodes.
  - The event *Jesus's appearance before Herod*, unique to Luke's Gospel, generates only a single node.

  This results in a graph with a total number of nodes equal to the sum of all available event versions across the documents, not simply 4 (gospels) $\times$ 169 (events).
- **Edges:** The graph contains two functionally distinct types of edges:
  1. **Temporal Edges (`BEFORE`):** Directed edges connect nodes that are sequential within the same Gospel, enforcing the intra-document narrative flow.
  2. **Anchoral Edges (`SAME_EVENT`):** Undirected edges interconnect all nodes (versions) that refer to the same canonical event, creating localized clusters for version selection.

Figure 1 illustrates a fragment of the TAEG built from our data, showing three canonical events with varying numbers of version nodes, the directed `BEFORE` edges within each Gospel, and the `SAME_EVENT` clusters.

The rationale for including *all* available versions of each event in the graph, rather than pre-selecting one, is threefold. First, the `SAME_EVENT` cluster defines the candidate space for version selection: centrality is meaningful precisely because all competing versions are interconnected and can be compared within the local context of the event. Second, retaining every version preserves the complementary details required by the Completeness and Representativeness objectives, instead of discarding information at graph-construction time. Third, the clusters provide the structural locus for future extensions, such as contradiction handling (Section 3.3) and abstractive fusion of the versions of an event.

4.2.2 Consolidation Algorithm

With the TAEG constructed, the consolidation process becomes a deterministic algorithm guided by the graph's structure, as detailed in Algorithm 1. The LexRank algorithm is applied to this structure, where its centrality score is repurposed to evaluate which version of an event is most representative within its local context. This transforms LexRank from a generic sentence selector into a powerful, temporally-aware version selection mechanism.

We emphasize that the resulting consolidation is purely *extractive*: for each canonical event, the algorithm outputs verbatim the full text of the selected version, and no sentence is rewritten or fused. The Representativeness objective in its strongest form — an enriched version combining all descriptions of an event — therefore requires abstractive fusion, which the TAEG enables structurally (the `SAME_EVENT` clusters delimit exactly what should be fused) and which is investigated in a companion study by the authors [Finger *et al.*, 2026].

4.3 Timeline-Aware Reference Baselines

Since the TAEG leverages the canonical timeline as an architectural prior, a comparison restricted to the timeline-agnostic LexRank baseline would conflate two factors: the availability of the temporal prior and the graph-based selection mechanism. To disentangle them, we define a set of heuristic baselines that receive *exactly the same* canonical timeline as the TAEG. All of them follow the structure of Algorithm 1, iterating over the timeline T and emitting one version per event; they differ only in the selection criterion replacing the centrality-based choice (line 8):

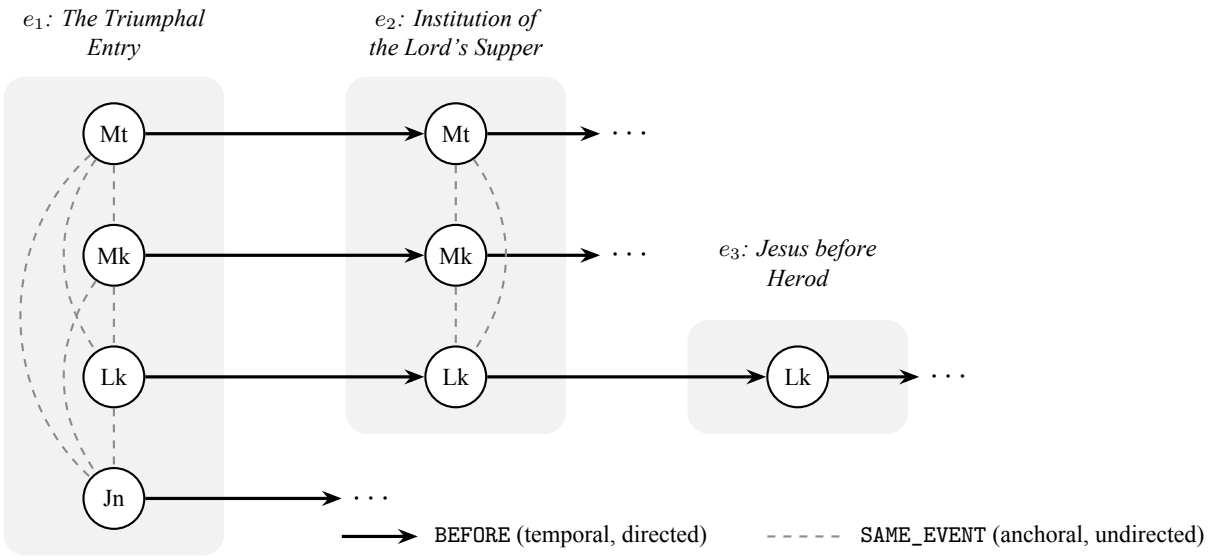

Figure 1. An illustrative fragment of the TAEG for three canonical events of the Holy Week timeline, in chronological order. Each circle is a version node, i.e., the description of a canonical event in one Gospel (Mt: Matthew; Mk: Mark; Lk: Luke; Jn: John). Solid directed edges are temporal `BEFORE` edges, linking sequential events within the same Gospel; dashed undirected edges are anchoral `SAME_EVENT` edges, interconnecting all versions of the same canonical event into a localized cluster. The number of nodes per event varies with the number of Gospels that describe it (four, three, and one, respectively).

Algorithm 1 TAEG-based Narrative Consolidation

function ConsolidateNarrative(D, T)
Require: Set of documents D, Canonical timeline T
Ensure: A single consolidated narrative text
    $G \leftarrow \text{BuildTAEG}(D, T)$ ▷ Build the graph
    $Scores \leftarrow \text{RunLexRank}(G)$ ▷ Compute centrality
    $OutputText \leftarrow$ empty string
    for each event e_j in timeline T **do**
        $CandNodes \leftarrow \text{GetEventNodes}(G, e_j)$
        if $CandNodes$ is not empty **then**
            $BestNode \leftarrow \text{FindMaxScore}(CandNodes, Scores)$
            Append(OutputText, BestNode.text)
        end if
    end for
    return $OutputText$
end function

- **Timeline+Random:** selects, for each event, one of the available versions uniformly at random. Results are averaged over 30 independent runs (mean ± standard deviation). This baseline establishes the performance floor attainable from the temporal structure alone.
- **Timeline+Priority:** selects the version according to a fixed source-priority order (Matthew ≻ Mark ≻ Luke ≻ John), reflecting a simple editorial policy.
- **Timeline+Centroid:** selects, for each event, the version with the highest mean TF-IDF cosine similarity to the other versions of the same event — a purely *local* selection that uses no global graph information.
- **Timeline+Longest:** selects, for each event, the longest available version (in tokens). This is a strong naive heuristic for the task, since the reference consolidation fuses details from all sources and longer versions tend to be more complete.

By construction, every timeline-aware method — including the TAEG — achieves perfect temporal coherence; the comparison among them is therefore decided purely by content quality, isolating the contribution of the graph structure and centrality-based version selection. Together with the timeline-agnostic LexRank, these methods form a graded ladder of reference baselines for the Narrative Consolidation task, which we expect to be useful for benchmarking future approaches. All implementations are available in our public repository.

5 The Gospel Consolidation Language Resource

Our evaluation dataset represents a contribution to the field, providing a standardized resource for narrative consolidation research. This resource will be made publicly available to ensure reproducibility and encourage further research.

5.1 Source Texts and Temporal Structure

The dataset focuses on the Holy Week period from the four Gospels (Matthew, Mark, Luke, and John) in the English New International Version [Biblica, 2011]. This narratively dense segment, while representing only a single week, constitutes approximately 35% of the total Gospels text, making it an ideal corpus for studying multi-perspective accounts. We adapted the chronological framework from Aschmann's Gospel harmony [Aschmann, 2022], which identifies 169 canonical events during the Holy Week. This framework provides the external temporal knowledge assumed as input by the task while representing scholarly consensus on event ordering.

5.2 Golden Sample Reference

As ground truth for evaluation, we used a manually created, high-quality reference consolidation (Golden Sample) that represents an ideal fusion of the Gospel narratives. This

reference maintains chronological order while incorporating the most significant details from each source into a fluid narrative. The methodology and the resource itself were first presented at the IV Congresso Brasileiro de Humanismo Solidário na Ciência [Cunha, 2025].

5.3 Public Availability and Format

The complete dataset, including processed texts, temporal alignments, and evaluation resources, is publicly available at `https://github.com/neemias8/TAEG`. While this system was validated using the English NIV of the Gospels, the "book:chapter:verse" system makes the framework applicable to any other language and/or version of them.

6 Experimental Setup

We designed our evaluation as a comparison along a graded ladder of methods: the traditional multi-document summarization baseline (timeline-agnostic LexRank), the timeline-aware heuristic baselines of Section 4.3, ablated variants of the TAEG, and the full TAEG-based narrative consolidation approach.

6.1 Evaluation Metrics

Evaluating the quality of multi-document summaries (MDS) is a challenging task. As pointed out by [Ma *et al*., 2023], standard evaluation metrics, while widely used, are not ideal for the MDS task, with notable criticisms regarding the lack of semantic awareness in metrics like ROUGE. However, in the absence of a superior, universally accepted evaluation method, we employ a suite of well-established metrics to assess different aspects of the generated summaries. This section consolidates the operational definitions of all metrics used in this study (cf. Section 3.4).

6.1.1 Semantic Quality Metrics

- **ROUGE** [Lin, 2004]: We compute ROUGE-1, ROUGE-2, and ROUGE-L F1 scores to measure lexical overlap with the Golden Sample. ROUGE-L, in particular, measures the longest common subsequence and thus intrinsically rewards the preservation of correct sequential order.
- **METEOR** [Banerjee and Lavie, 2005]: Provides word alignment-based evaluation with consideration for synonymy and stemming.
- **BERTScore** [Zhang *et al*., 2020b]: Measures semantic similarity using BERT embeddings, capturing deeper semantic relationships.

We note that ROUGE and BERTScore are conceptually similar — both assess content similarity between the generated text and the reference — differing in the representation level at which matching is performed: surface n-grams in ROUGE versus contextualized embeddings in BERTScore, the latter being more robust to paraphrasing.

6.1.2 Temporal Coherence Metric

Kendall's Tau [Kendall, 1938]: This rank correlation coefficient measures the agreement between the sentence order in the generated narrative and the canonical chronological order. A score of 1.0 indicates perfect temporal agreement, while scores closer to 0 indicate temporal disorder. This metric is central to our evaluation as it directly measures the primary objective of narrative version consolidation. We emphasize that any method that iterates over the canonical timeline — the TAEG and all baselines of Section 4.3 — achieves $\tau = 1.000$ by construction; for these methods, the metric is verified directly on the emitted sequence of event identifiers and reported as a sanity check. The heuristic sentence-to-event matcher used to estimate τ for arbitrary text is applied only to the timeline-agnostic LexRank baseline, since its estimate is sensitive to the per-event version choice and would conflate version selection with ordering for timeline-aware outputs.

6.1.3 Selection-Level Evaluation

Corpus-level content metrics are diluted on this task: 72 of the 169 events have a single available version (identical output for every timeline-aware strategy), parallel accounts of the same event are mutually similar, and BERTScore saturates. We therefore additionally evaluate the version-selection component directly. For each *contested* event (two or more available versions), we define the *oracle* as the version with the highest ROUGE-L F1 against that event's segment of the Golden Sample (which is aligned to the canonical events), and report each strategy's **oracle selection accuracy**: the fraction of contested events in which it picks the oracle. The analytical expectation for uniform random selection on this dataset is ≈ 0.35. We propose this selection-level evaluation as the appropriate instrument for the extractive setting of Narrative Consolidation, complementing the corpus-level proxies.

7 Benchmark Results

This section reports the behaviour of the reference systems of Section 4 on the Gospel Consolidation Language Resource. The presentation follows the two factors of the ladder: we first quantify what access to a canonical timeline buys (Sections 7.1–7.4), then compare the version-selection criteria under equal temporal information (Sections 7.5–7.8). Section 8 synthesizes the findings and their implications for the task.

7.1 The Effect of the Temporal Backbone

Table 2 contrasts the timeline-agnostic LexRank reference with the TAEG, taken here as a representative timeline-aware system, both evaluated against the Golden Sample. The comparison quantifies the combined effect of adding the temporal backbone and a structured selection mechanism; Section 7.5 then separates the two.

Metric	Baseline	TAEG	Improv.
ROUGE-1 F1	0.887	**0.918**	+3.5%
ROUGE-2 F1	0.712	**0.848**	+19.1%
ROUGE-L F1	0.206	**0.846**	+310.7%
BERTScore F1	0.835	**0.906**	+8.5%
METEOR	0.453	**0.550**	+21.4%
Kendall's Tau	0.320	**1.000**	by design
Length (chars)	81,418	74,280	-

Table 2. Effect of the temporal backbone: the timeline-agnostic LexRank reference versus the TAEG, a representative timeline-aware system. Perfect temporal ordering is an architectural property of every timeline-aware system (by design), not an empirical finding; see Section 7.2. The full comparison among timeline-aware systems is given in Table 4.

7.2 Temporal Coherence Analysis

Every timeline-aware system attains $\tau = 1.000$ by construction: iterating over a pre-ordered list of canonical events forces the output into the correct chronological sequence, so the value is reported as a verified sanity check rather than as an empirical gain (Section 6.1). The property matters because it decouples chronological structuring from content selection, leaving the selection criterion as the only free variable (Section 7.5). In contrast, the timeline-agnostic LexRank reference produces a narrative with significant temporal disorder (Tau = 0.320), confirming that a graph based on semantic similarity cannot preserve narrative flow.

7.3 Content Quality

Every content metric improves relative to the timeline-agnostic reference. The gain in ROUGE-L F1 (+310.7%) follows directly from the guaranteed chronological integrity, since ROUGE-L measures the longest common subsequence and a correctly ordered output holds a decisive structural advantage over a disordered one; the smaller gains in ROUGE-1 (+3.5%), ROUGE-2 (+19.1%) and METEOR (+21.4%) reflect the version-selection component on top of the temporal structure. BERTScore rises from 0.835 to 0.906, but saturates on this task — remaining high for any reasonably complete extractive output — and is therefore the least discriminative of our metrics (cf. Section 7.8).

7.4 Conciseness vs. Consolidation Analysis

The results presented for the standard LexRank baseline in Table 2 reflect a parameter setting of 750 sentences. To further investigate the trade-off between conciseness and narrative quality, we analyzed the baseline's performance across various summary lengths. Table 3 shows the results for summaries constrained to different sentence counts, contrasting them with the timeline-aware consolidation.

The analysis in Table 3 clearly demonstrates that simply increasing the number of sentences in the summary does not address the fundamental problem of narrative coherence. While some metrics like ROUGE-1 and METEOR improve up to a certain point before declining, the temporal coherence (Kendall's Tau) remains consistently low, indicating a persistently disordered narrative. Even at its peak performance, the standard LexRank approach fails to come close to the quality and temporal integrity of the TAEG-based method. This

	LexRank Baseline				TAEG
Sentences	**100**	**500**	**1000**	**1500**	**N/A**
ROUGE-1 F1	0.296	0.804	0.862	0.784	**0.918**
ROUGE-2 F1	0.263	0.655	0.728	0.733	**0.848**
ROUGE-L F1	0.129	0.206	0.199	0.188	**0.846**
BERTScore F1	0.835	0.835	0.835	0.835	**0.906**
METEOR	0.097	0.361	0.483	0.484	**0.550**
Kendall's Tau	0.268	0.305	0.320	0.320	**1.000**
Length (chars)	15K	59K	101K	129K	**74K**

Table 3. Conciseness vs. Consolidation Analysis. Comparison of metrics between different LexRank (Baseline) summary lengths against the consolidated TAEG output.

experiment reinforces our central argument: for long and complex narratives, comprehensive coverage and chronological soundness are far more critical than mere conciseness.

7.5 Comparison of Reference Systems under Equal Temporal Information

Table 4 is the central comparison of this study. It places all timeline-aware reference systems of Section 4 side by side; since every one of them receives the same canonical timeline and achieves $\tau = 1.000$ by construction, the observed differences are attributable solely to the version-selection criterion. The comparison therefore measures how much content-selection sophistication is worth on this task once chronology is given.

Metric	T+Rand.	T+Prior.	T+Centr.	T+Long.	TAEG
ROUGE-1 F1	0.889 $\pm$.009	0.886	0.891	**0.958**	0.918
ROUGE-2 F1	0.816 $\pm$.013	0.814	0.821	**0.938**	0.848
ROUGE-L F1	0.813 $\pm$.015	0.811	0.818	**0.947**	0.846
BERTScore F1	0.932 $\pm$.033	0.897	0.935	**0.995**	0.906
METEOR	0.478 $\pm$.026	0.453	0.472	**0.639**	0.550
Kendall's Tau	1.000	1.000	1.000	1.000	1.000
Length (chars)	70.1K	69.3K	69.9K	79.2K	74.3K

Table 4. Timeline-aware comparison: all methods iterate over the same canonical timeline and differ only in the per-event version-selection criterion. Timeline+Random is reported as mean $\pm$ std over 30 runs. Kendall's Tau is 1.000 by design for every method in this table (verified on the emitted event sequence).

The comparison yields two results. First, the spread among timeline-aware systems (ROUGE-L F1 from 0.811 to 0.947) is an order of magnitude smaller than the gap between them and the timeline-agnostic reference (0.206). Even uniform random selection reaches 0.813: once the chronological structure is given, a large share of the task is already solved, and the choice of version accounts for the remainder. Second, within that remainder, the ordering of the systems is instructive. The centrality-based selection over the TAEG (0.846) is ahead of random (0.813), fixed-priority (0.811), and local-centroid (0.818) selection, but behind the *Timeline+Longest* heuristic (0.947), which leads on every content metric.

The advantage of the length heuristic follows from how the reference is built. The Golden Sample is a *fusion* of all sources, so the most detailed account of an event overlaps it most; length is thus a close proxy for the oracle choice on this benchmark (Section 7.8), and the property should be expected on any fusion-style reference over complementary, non-conflicting sources. It does not transfer to corpora where

accounts diverge or contradict one another (Section 3.3), since there the longest account is not necessarily the most faithful one. A second limit is intrinsic to the extractive setting: selecting one version per event recovers at best the richest single account, never the union of details spread across sources, which Section 3.1 identifies as the ideal output under the Representativeness objective. The challenge posed by this resource therefore has two parts: *surpassing the length heuristic with a principled selection mechanism*, and *moving beyond selection to the fusion of complementary details* — the latter pursued in a companion study [Finger *et al*., 2026].

For full transparency, we note that the configuration reported in the originally submitted version of this manuscript corresponds to the Timeline+Longest column of Table 4; the implementation of Algorithm 1 was completed and audited during the first revision, and every configuration is now reported under its correct label.

7.6 Ablation Study

To quantify the contribution of each edge type of the TAEG, we evaluate two ablated variants: (i) *TAEG w/o* `BEFORE`, in which temporal edges are removed from the graph while the final loop still iterates over the timeline; and (ii) *TAEG w/o* `SAME_EVENT`, in which anchoral edges are removed, assessing the impact of event grouping on version selection. Table 5 presents the results.

Metric	w/o BEFORE	w/o SAME_EV.	TAEG
ROUGE-1 F1	0.886	0.929	0.918
ROUGE-2 F1	0.811	0.865	0.848
ROUGE-L F1	0.809	0.866	0.846
BERTScore F1	0.935	0.964	0.906
METEOR	0.469	0.575	0.550
Selection acc.	0.341	0.489	0.489

Table 5. Ablation of the TAEG's edge types. Both variants keep the final timeline iteration, so all configurations preserve $\tau = 1.000$ by design. Selection accuracy is the oracle selection accuracy of Section 7.8 (random floor ≈ 0.37).

Before interpreting the results, it is important to delimit what each ablation removes. Both variants remove *edges from the graph over which centrality is computed*; neither removes the event grouping itself, which is not a component of the TAEG but part of the task input — the alignment between event versions and canonical events comes from the chronology, and it is what defines the candidate set of each event in line 6 of Algorithm 1. Removing the grouping would not ablate a design choice: with no candidate set, the selection step would maximize a fixed global score over the whole graph and emit the same version at every position, producing a degenerate output. The ablations therefore answer a well-posed question — what does each relation contribute *as weighted evidence for centrality* — and not whether event clustering is needed.

With that scope in mind, the results give a clear and partly unexpected picture. Removing the `BEFORE` edges collapses selection quality to the random floor (oracle accuracy 0.341; ROUGE-L 0.809, indistinguishable from Timeline+Random): the discriminative signal exploited by centrality flows through the narrative-flow similarities encoded by the temporal edges. Removing the `SAME_EVENT` edges, in contrast, leaves oracle accuracy unchanged at 0.489 and slightly improves the corpus metrics. The coincidence of the accuracy figures does not mean that the two configurations behave identically: their outputs differ, as the divergent corpus metrics and output lengths in Table 5 show, so a number of per-event choices do change — they simply trade correct and incorrect selections in equal measure.

The explanation is that parallel accounts of the same event are nearly equidistant in the TF-IDF space: any version resembles its siblings about as much as any other, so intra-cluster weights are close to uniform and carry almost no information for ranking. This is a negative result, and we report it as such, because it is informative for the design of future systems: it indicates that a graph over event versions should not expect intra-event lexical similarity to discriminate between candidates, and that models building on this structure — including graph neural approaches — will need weights derived from semantic, discourse, or factuality features rather than from surface overlap. The corollary for the TAEG is equally direct: of its two relations, only the temporal one earns its place as evidence for selection.

7.7 Sensitivity to Incomplete Timelines

Real-world chronologies are rarely complete. To provide a first quantification of the framework's robustness to incomplete temporal knowledge, we randomly remove a percentage of the canonical events from the timeline before building the TAEG, and measure the impact on all metrics. Table 6 reports the results, averaged over 10 runs per degradation level.

Metric	0%	10%	25%	50%
ROUGE-L F1	0.846	0.795 $\pm$.022	0.723 $\pm$.037	0.534 $\pm$.040
BERTScore F1	0.906	0.897 $\pm$.015	0.882 $\pm$.021	0.878 $\pm$.033
METEOR	0.550	0.474 $\pm$.021	0.397 $\pm$.027	0.246 $\pm$.026

Table 6. Impact of randomly removing canonical events from the timeline (degradation level in % of events removed; mean $\pm$ std over 10 runs). Removed events are absent from the output; among the remaining events, ordering stays correct by design.

Content metrics degrade gracefully and roughly in proportion to the lost coverage, confirming that the impact of an incomplete timeline is completeness-driven rather than disruptive: among the remaining events, ordering and selection are unaffected. Notably, even with half of the timeline removed, the consolidation retains a ROUGE-L F1 of 0.534 — still far above the timeline-agnostic LexRank baseline (0.206) operating with full access to the texts. A partial temporal backbone is thus considerably better than none, a result that establishes the robustness reference that automatic (and hence imperfect) event-alignment methods will have to meet.

7.8 Selection-Level Evaluation

Table 7 reports the oracle selection accuracy of Section 6.1.3, computed over the 88 contested events for which a golden segment could be reliably aligned (out of 96 contested events; the analytical random floor on this set is ≈ 0.35, empirically 0.380). This selection-level view discriminates the strategies far more sharply than the (saturating) corpus metrics.

Strategy	Oracle selection accuracy
Timeline+Random	0.372 $\pm$ 0.052
Timeline+Priority	0.409
Timeline+Centroid	0.375
Timeline+Longest	**0.648**
TAEG (Algorithm 1)	0.489
TAEG w/o `BEFORE`	0.341
TAEG w/o `SAME_EVENT`	0.489

Table 7. Oracle selection accuracy on the 88 evaluated contested events (oracle = version with highest ROUGE-L F1 against the event's golden segment; random floor ≈ 0.37).

Three observations follow. First, the TAEG's centrality-based selection (0.489) is clearly and significantly above the random floor: across 30 seeded random runs, it sits at the 100th percentile of the random distribution for oracle accuracy, ROUGE-1/2/L, and METEOR — the graph carries a genuine selection signal. Second, the purely local Timeline+Centroid baseline performs at the floor (0.375), showing that intra-cluster similarity alone is uninformative (parallel accounts are nearly equidistant) and that the TAEG's advantage over it comes from the global graph structure. Third, Timeline+Longest reaches 0.648: with a fusion-style reference over complementary sources, the most detailed version is most often the best available one, making length a remarkably strong proxy *on this benchmark*. We report this candidly: it characterizes the benchmark as much as the methods, and it sharpens the research question for noisier domains, where detail and faithfulness no longer coincide.

8 Findings and Implications for the Task

Benchmarking the reference systems produces four findings. Because this study aims to establish the task rather than to advocate a method, we state them as properties of Narrative Consolidation and of the resource, and draw out what each implies for future work.

Finding 1: chronological structure, not content selection, is the defining difficulty. The gap between timeline-agnostic and timeline-aware systems (ROUGE-L F1 0.206 vs. 0.811–0.947) dwarfs the spread among selection criteria. This is the empirical justification for treating Narrative Consolidation as a task distinct from Multi-Document Summarization: a strong extractive summarizer applied without chronological structure does not merely score lower, it produces an output of a different kind, in which the sequence of events is lost (Section 4.1). It also implies that progress on this task will come primarily from how chronology is obtained and represented — and, since the timeline is an input here, that the induction of chronological structure is the critical open problem (Section 9).

Finding 2: a length heuristic is a strong reference point on fusion-style references. Selecting the longest available version per event attains 0.947 ROUGE-L F1 and 0.648 oracle selection accuracy, ahead of every other system tested. Simple heuristics that resist more elaborate alternatives are a familiar phenomenon in summarization, where lead-based baselines long remained competitive with sophisticated models; the analogous observation for this task is that when the reference consolidation fuses complementary, non-conflicting accounts, the detail of an account and its quality nearly coincide. Two implications follow. For evaluation: any future system on this resource must be compared against a length heuristic, not only against a random or centrality-based one, and we supply it for that purpose. For resource design: benchmarks in which the sources genuinely conflict are needed to separate detail from faithfulness, which is why we identify a news-domain resource as a priority (Section 9).

Finding 3: an explicit relational prior yields signal, but not enough. Centrality over the TAEG selects the oracle version in 48.9% of contested events, against a random floor of about 37% and at the 100th percentile of the seeded random distribution — the structure carries genuine information — yet it remains below the length heuristic. Structural priors, therefore, are not automatically superior to trivial ones on this task; the burden on a structured method is not to beat chance but to beat completeness. We state this as the concrete open challenge posed by the benchmark.

Finding 4: intra-event lexical similarity is uninformative for version selection. The ablation of Section 7.6 shows that the discriminative signal lies in the temporal relations, while similarity weights among versions of the same event are effectively uniform. For future models — in particular graph neural architectures over this structure — the implication is that useful intra-event edge weights must encode semantic, discourse, or factuality distinctions rather than surface overlap. Reporting this negative result is, in our view, one of the more useful outcomes of the benchmark, since it removes a design avenue that would otherwise appear natural.

9 Limitations and Future Work

While our results are promising, we acknowledge several limitations, which we present as an explicit research roadmap.

First, this study assumes an external, pre-defined timeline. As discussed in Section 3.1, this was a deliberate scoping decision that isolates the version-selection subproblem under a known chronology. Automatically inducing the timeline from the documents themselves requires solving cross-document event extraction, cross-document event coreference, and temporal ordering — each an open research problem in its own right — and is precisely the focus of the authors' ongoing work, as the next stage of this research program. The sensitivity analysis of Section 7.7 provides the robustness reference that such automatic (and hence imperfect) alignment methods will have to meet. Handling *contradictory* (rather than merely incomplete) timelines remains open.

Second, our evaluation is based on a single, albeit complex, dataset from the religious domain, which, as discussed in Section 3.3, deliberately controls for factual contradictions and adversarial perspectives. Testing the task and its reference systems on other multi-perspective narrative corpora, such as legal depositions or journalistic reports, is crucial to assess its generalizability. To the best of our knowledge, however, no existing corpus simultaneously provides multiple overlapping accounts, an aligned canonical timeline, and a reference consolidation; constructing such a benchmark for the news domain — including annotations of contradictions

and perspectives — is a substantial research contribution that we leave as future work.

Third, regarding stronger baselines from the Timeline Summarization literature: off-the-shelf TLS/MDTS systems are date-driven, relying on explicit temporal expressions and document timestamps to anchor content, and producing lists of dated summaries rather than a single continuous narrative. Since our corpus contains no explicit dates and the task's target output is a fluid narrative, a direct comparison would be ill-defined. The timeline-aware baselines of Section 4.3 play this role instead, granting competing methods the same temporal knowledge in a format compatible with the task; adapting TLS methods to Narrative Consolidation remains an interesting direction.

Finally, the natural extension of this work is abstractive: leveraging the TAEG as a structural backbone — e.g., with a GNN encoder [Yasunaga *et al*., 2017] and an abstractive decoder [Xiao *et al*., 2022; Zhang *et al*., 2020a] — so that the versions within each `SAME_EVENT` cluster are fused into enriched event descriptions. A first investigation of abstractive Narrative Consolidation on this same resource is presented in a companion study by the authors [Finger *et al*., 2026].

9.1 Influence of Genre, Domain, and Language

Since the present findings rest on a single resource, we note how they depend on its properties. **Genre** governs the relationship between detail and faithfulness, and therefore the strength of the length heuristic (Finding 2): the Gospels are complementary testimonial accounts with stable content and no adversarial intent, so a longer account is usually a more informative one, whereas in breaking news, where successive reports revise earlier facts, and in legal depositions, where accounts are strategically framed, we would expect detail to decouple from faithfulness, increasing the value of selection mechanisms that model contradiction and stance (Section 3.3). Genre also determines what temporal evidence is available: date-wise timeline summarization presupposes explicit temporal expressions and document timestamps [Martschat and Markert, 2017; Mamidala and Sanampudi, 2021], and dedicated temporal reasoning is required where such signals are implicit or noisy [Song *et al*., 2025], whereas the narrative testimony studied here supplies neither. **Domain** determines whether a canonical timeline exists at all: our assumption is realistic for domains with an established chronology — scriptural harmonies, case files, clinical histories — but not for open-domain collections, where the chronology must be induced, and the degradation analysis of Section 7.7 covers missing events but not *erroneous* ordering, the more likely failure mode of an induced timeline. **Language** affects the resource and the metrics differently. The resource is largely language-independent, since the alignment is expressed over `book:chapter:verse` references, so the same 169-event chronology applies to any translation and parallel benchmarks can be built at low cost. The metrics are not: ROUGE matches surface n-grams [Lin, 2004], so morphological variation understates agreement; METEOR depends on stemming and synonymy modules that must be available for the target language [Banerjee and Lavie, 2005]; and BERTScore inherits the quality of the contextual embeddings trained for it [Zhang *et al*., 2020b]. Establishing how these proxies behave across languages is thus a prerequisite for cross-lingual replication, which we have not undertaken.

10 Conclusion

This paper introduced and formalized *Narrative Consolidation*, a new NLP task that reframes the processing of multiple overlapping narratives, arguing for the primacy of coherence and completeness over simple compression. Around that definition we assembled what the task needs in order to be studied: an evaluation paradigm that includes a selection-level metric, the Gospel Consolidation Language Resource, and a graded suite of reference systems spanning timeline-agnostic centrality, timeline-aware heuristics, and the Temporal Alignment Event Graph, together with ablations and a robustness analysis.

Benchmarking these systems characterizes the task more sharply than any of them individually. Chronological structure, rather than content-selection sophistication, is the dominant factor: granting a canonical timeline moves ROUGE-L F1 from 0.206 to at least 0.811, while the criterion used to choose among versions accounts for the remaining margin. Within that margin, an explicit relational structure yields a real but modest signal, and is outperformed on this resource by a simple length heuristic — an outcome that says as much about fusion-style references as about the systems, and that we therefore record as a property of the benchmark. The ablations add a negative result of direct use to future designs: the discriminative evidence lies in the temporal relations, whereas lexical similarity among versions of the same event is uninformative.

We consequently offer the TAEG not as a solution to Narrative Consolidation but as one reference point among several, and we leave a concrete challenge in its place: on this benchmark, no principled selection mechanism has yet surpassed selecting the longest account; doing so — particularly on resources where sources genuinely conflict — and moving beyond selection to the fusion of complementary details are the natural next objectives. Alongside it, the induction of the chronological backbone without an external timeline remains, in our assessment, the decisive open problem for the task. The resource, the reference implementations, and the evaluation code are publicly available so that both challenges can be taken up directly.

Declarations

Acknowledgements

The authors thank the anonymous reviewers for their constructive feedback.

Authors' Contributions

RF is the main contributor and writer of this manuscript. EC and GR reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially supported by CNPq (grant 313845/2023-9).

Availability of data and materials

The datasets and software generated and analysed during the current study are available at https://github.com/neemias8/TAEG.

References

Aschmann, R. (2022). Chronology of the four gospels (harmony of the gospels). Available at:https://www.biblechronology.net/ChronologyOfTheFourGospels.pdf.

Banerjee, S. and Lavie, A. (2005). METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and Summarization*, pages 65–72. Available at:https://aclanthology.org/W05-0909/.

Barzilay, R., McKeown, K. R., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proc. Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 550–557. DOI: 10.3115/1034678.1034762.

Biblica, I. (2011). *The Holy Bible, New International Version*. Book. Copyright © 1973, 1978, 1984, 2011 by Biblica, Inc.™ Used by permission for academic research.

Bollegala, D., Okazaki, N., and Ishizuka, M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Inf. Process. Manage.*, 46(1):89–109. DOI: 10.1016/j.ipm.2009.07.004.

Cunha, A. (2025). Semana da paixão unificada: uma harmonização moderna dos evangelhos. In *Anais do IV Congr. Bras. de Humanismo Solidário na Ciência*. SBCC. Book.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479. DOI: 10.1613/jair.1523.

Eusebius (1999). *The Church History*. Kregel Publications. Book.

Finger, R. A., Cortes, E. G., Rigo, S. J., and Ramos, G. d. O. (2026). Neurosymbolic narrative consolidation: Grounding abstractive MDS and LLMs with temporal event graphs. In *Proc. IEEE Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE. Available at:https://linklings.s3.amazonaws.com/organizations/WCCI/wcci2026/submissions/stype114/LpSsT-ijcnn_pap1093s2.pdf.

Jiao, P., Chen, H., Guo, X., Zhao, Z., He, D., and Jin, D. (2025). A survey on temporal interaction graph representation learning: Progress, challenges, and opportunities. DOI: 10.48550/arXiv.2505.04461.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93. DOI: 10.1093/biomet/30.1-2.81.

Lin, C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries. In *Proc. Workshop on Text Summarization Branches Out*, pages 74–81. Available at:https://aclanthology.org/W04-1013/.

Ma, C., Zhang, W. E., Guo, M., Wang, H., and Sheng, Q. Z. (2023). Multi-document summarization via deep learning techniques: A survey. *ACM Comput. Surv.*, 55(5):1–37. DOI: 10.1145/3529754.

Mamidala, K. K. and Sanampudi, S. K. (2021). A novel framework for multi-document temporal summarization (MDTS). *Emerg. Sci. J.*, 5(2):184–190. DOI: 10.28991/esj-2021-01268.

Martschat, S. and Markert, K. (2017). Improving ROUGE for timeline summarization. In *Proc. Conf. of the European Chapter of the ACL (EACL)*, pages 285–290. DOI: 10.18653/v1/E17-2046.

Petersen, W. L. (1994). *Tatian's Diatessaron: Its Creation, Dissemination, Significance, and History in Scholarship*. Brill. DOI: 10.2307/3266926.

Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., and Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. DOI: 10.48550/arXiv.2006.10637.

Santana, B., Campos, R., Amorim, E., Jorge, A., Silvano, P., and Nunes, S. (2023). A survey on narrative extraction from textual data. *Artif. Intell. Rev.*, 56(8):8393–8435. DOI: 10.1007/s10462-022-10338-7.

Song, J., Akhter, M. E., Atzil-Slonim, D., and Liakata, M. (2025). Temporal reasoning for timeline summarisation in social media. In *Proc. Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 28085–28101. DOI: 10.18653/v1/2025.acl-long.1362.

Xiao, W., Beltagy, I., Carenini, G., and Cohan, A. (2022). PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *Proc. Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 4016–4036. DOI: 10.18653/v1/2022.acl-long.276.

Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., and Radev, D. (2017). Graph-based neural multi-document summarization. In *Proc. Conf. on Computational Natural Language Learning (CoNLL)*, pages 452–462. DOI: 10.18653/v1/K17-1045.

Yu, Y., Jatowt, A., Doucet, A., Sugiyama, K., and Yoshikawa, M. (2021). Multi-TimeLine summarization (MTLS): Improving timeline summarization by generating multiple summaries. In *Proc. Annual Meeting of the Assoc. for Computational Linguistics and Int. Joint Conf. on NLP (ACL-IJCNLP)*, pages 377–387. DOI: 10.18653/v1/2021.acl-long.32.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020a). PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 11328–11339. PMLR. DOI: 10.48550/arXiv.1912.08777.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). BERTScore: evaluating text generation with

BERT. In *Proc. Int. Conf. on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.1904.09675.