IIITH Boys at SemEval-2026 Task 4: StoryNet - Understanding Narrative Story Similarity through Symbolic Representations

Amol Vijayachandran*, **Ananth Rajesh***, **Siddharth Mago***,
Maitreya Prafulla Chitale, **Aparajitha Allamraju**,

IIIT Hyderabad
{amol.vijayachandran, ananth.rajesh, siddharth.mago}@students.iiit.ac.in
{maitreya.chitale, aparajitha.allamraju}@research.iiit.ac.in

Abstract

Narrative similarity extends beyond standard semantic tasks, requiring alignment of temporal, causal, and emotional structures. We present StoryNet, a framework that represents stories as heterogeneous graphs with character, event, and theme nodes. Stories are decomposed into structured narrative facets using large language models, and similarity is evaluated through both weighted semantic facet comparison and a graph neural network trained with contrastive learning. We analyze how integrating symbolic structure with learned graph representations compares to purely embedding-based baselines. We release our code[1] and resources[2] publicly.

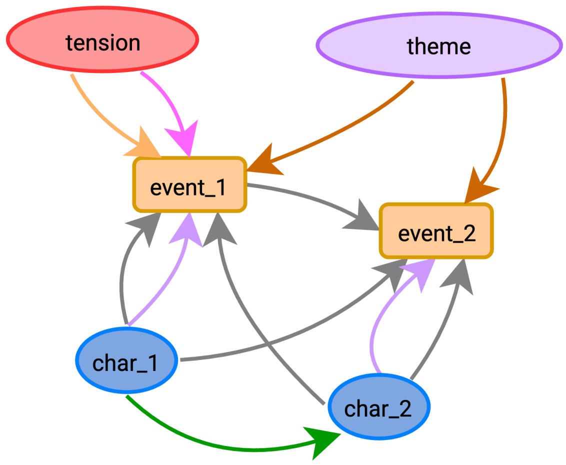


Figure 1: Example heterogeneous story graph showing character, event, and theme nodes with their relational edges.

1 Introduction

Narrative similarity has traditionally been modeled through lexical overlap or distributed embeddings (Mihalcea and Csomai, 2006; Reimers and Gurevych, 2019), with pretrained language models achieving strong performance on semantic comparison tasks (Devlin et al., 2019). However, these approaches encode narrative structure implicitly, leaving elements such as character relations, causal chains, and thematic progression without explicit representation. Structured models offer a complementary perspective by directly encoding entities and their dependencies.

We introduce StoryNet, a graph-based framework that represents each story as a heterogeneous graph of *characters*, *events*, and *themes* with their relational dependencies. Within this framework, we investigate two complementary approaches: (i) a facet-based semantic similarity method that compares stories across thematic, emotional, causal, and relational dimensions with learned aggregation weights, and (ii) a Graph Neural Network trained with triplet-based contrastive learning to produce graph-level embeddings optimized for relative narrative similarity.

Together, these approaches position StoryNet as a unified framework for studying narrative similarity through explicit structure alongside learned graph representations. Our work investigates how structured narrative modeling can complement embedding-based methods for narrative comparison. An example of a constructed story graph is shown in Figure 1.

2 Related Work

Narrative and Semantic Similarity. Text similarity has traditionally been modeled using lexical overlap, corpus-based metrics, and distributed sentence embeddings (Mihalcea and Csomai, 2006; Reimers and Gurevych, 2019). With pretrained transformers (Devlin et al., 2019), contextual embeddings have become the dominant paradigm for semantic comparison. However, narrative similar-

*Equal contribution.
[1] https://github.com/AKR-END/StoryNet
[2] https://huggingface.co/AKRonHF/StoryNet

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3382–3393
July 3-4, 2026

ity extends beyond surface semantics, involving alignment in character roles, causal structure, and thematic development. Prior work on narrative modeling has explored event chains and script induction to capture recurring story structures (Chambers and Jurafsky, 2008; Pichotta and Mooney, 2016), underscoring the importance of structured representations for modeling plot progression.

Graph-Based Representations of Text and Narrative. Graphs have been used to represent linguistic and discourse structure, including dependency trees (Tesnière, 1959) and word co-occurrence networks (Mihalcea and Tarau, 2004). In longer documents, graph-based approaches model discourse and entity interactions to improve summarization and coherence (Zheng and Lapata, 2019; Wang et al., 2022). In narrative contexts, researchers have constructed character interaction networks (Agarwal et al., 2013; Elson et al., 2010), inferred interpersonal relationships (Srivastava et al., 2016), and built character-aware discourse graphs (Chitale et al., 2025). These works show that explicit relational modeling captures long-range dependencies that are difficult to encode in sequential architectures.

Graph Neural Networks and Representation Learning. Graph Neural Networks (GNNs), including Graph Convolutional Networks and Graph Attention Networks (Kipf and Welling, 2016; Veličković et al., 2018), provide a principled framework for learning representations from structured data. Combined with metric-learning objectives such as triplet loss (Schroff et al., 2015), GNNs produce embeddings aligned with similarity judgments. Building on these developments, our work constructs heterogeneous narrative graphs integrating characters, events, themes, and causal structure, and evaluates both symbolic similarity modeling and learned graph embeddings within a unified framework for narrative comparison.

3 Dataset Details

We use the official SemEval-2026 Narrative Similarity Task (Hatzel et al., 2026) dataset, a benchmark resource designed to evaluate models on cross-story similarity and representation learning. The dataset consists of narrative texts paired with human annotations, covering both relative similarity judgments and single-story embeddings.

For **Task A (Triplet Comparison)**, the dataset contains 2,139 annotated triplets for training and validation and 400 triplets for held-out testing. Each triplet comprises one anchor story and two candidate stories, with annotators indicating which candidate is more similar to the anchor. For model training, we combine the human-authored development subset with the synthetic subset released for the task, yielding a synthetic-heavy training mixture of 200 development examples and 1,897 synthetic examples. The official test set is released separately by the organizers.

For **Task B (Single-Story Embeddings)**, the same stories from Task A are reused. The goal is to encode each narrative into a vector representation suitable for retrieval and ranking. The evaluation methodology is not disclosed and is conducted entirely by the task organizers; we discuss our submitted results in Section 5.

4 System Overview

Our pipeline comprises three components: narrative facet extraction, graph construction and similarity computation.

4.1 Narrative Facet Extraction

Our methodology decomposes narrative analysis into three structured stages, each guided by carefully designed prompts to a large language model (LLM). Outputs are constrained to JSON schemas for downstream compatibility, with rule-based repair for malformed responses. The three stages are:

1. **Character & Relationship Extraction:** Identifies characters with their narrative roles, agency levels, and attributes, along with pairwise relationships annotated for type, evolution, and power dynamics.

2. **Event & Causality Extraction:** Extracts events with participants, emotional tone, narrative function, and temporal ordering, then infers causal chains and parallel narrative threads.

3. **Emotional & Thematic Analysis:** Captures emotional trajectories across characters, thematic elements at varying abstraction levels, and tension–resolution patterns including escalation, crisis, and resolution.

Full definitions of all categorical terms are provided in Appendix A, detailed stage descriptions in Appendix C, and extraction prompts in Appendix D.

4.2 Narrative Graph Construction

The extracted facets are assembled into a heterogeneous graph per story (see Figure 1 for an example). Nodes represent the three core narrative entities—characters, events, and themes—each enriched with the attributes produced by the corresponding extraction stage. Edges encode four relation types: character–character social ties, character–event participation links, directed event–event causal chains, and theme–event connections that reflect the pervasive influence of thematic elements across the narrative. This heterogeneous design preserves both the concrete dynamics of the plot and its higher-level abstractions in a single structure amenable to both symbolic comparison and neural message passing. Full node and edge specifications are provided in Appendix B.

4.3 Narrative Similarity via Graph Comparison

We evaluate narrative similarity with two complementary approaches: (i) a *semantic* approach that compares stories directly in facet space, aggregating facet scores with learned weights; and (ii) a *GNN* approach that learns graph-level embeddings end-to-end.

4.3.1 Semantic Similarity

The semantic approach models multiple narrative dimensions, including character roles and attributes, thematic abstractions, emotional arcs, causal structure, narrative functions, relationship dynamics, and event semantics. Each dimension is evaluated using a task-appropriate method, such as categorical overlap, distributional similarity, sequence alignment, or transformer-based embeddings (using `all-mpnet-base-v2`) for textual and thematic content.

These per-facet scores are combined using an 18-dimensional weight vector learned from Task A triplet supervision. The weights are optimized directly on the training mixture without explicit regularization, making them sensitive to correlations in the dominant synthetic subset of the data. As a result, facet importance learned during training does not consistently transfer under distribution shift to human-authored narratives.

The learned weight vector is further factorized into top-level facet weights and intra-facet component weights. This formulation enables interpretable comparison across narrative dimensions while capturing aspects such as thematic resonance and emotional development without relying on graph structure.

4.3.2 Learned Graph Embeddings (GNN)

While structural and semantic similarity can be measured independently, neither alone captures the full narrative fingerprint of a story. To learn unified representations that encode both graph topology and node-level semantics, we employ a Graph Neural Network (GNN) encoder trained in a triplet configuration.

Each story graph — whose nodes carry 768-dimensional semantic embeddings derived from textual attributes — is processed by a shared encoder that propagates information along the graph's edges via message passing. We adopt a Graph Attention Network (GAT) as the primary architecture, which applies multi-head self-attention during neighborhood aggregation. We hypothesize that this attention mechanism allows the model to learn to weight narratively salient connections — such as those involving pivotal characters or causally influential events — more heavily than peripheral ones, though verifying this via attention analysis remains future work. A GCN variant using spectral convolution-based aggregation was also evaluated; architectural details and comparisons for both are provided in Appendix E.

After message passing, node embeddings are aggregated via global mean pooling and linearly projected to produce a compact 1024-dimensional graph-level embedding. The model is trained using a triplet margin loss:

$$\mathcal{L}_{triplet} = \max(0,\ d(f(G_a), f(G_p)) \\ - d(f(G_a), f(G_n)) + \delta)$$

where $f(G)$ is the encoder output, subscripts a, p, n denote anchor, positive, and negative graphs, and δ is a fixed margin. This enforces a ranking constraint pulling similar story embeddings together while pushing dissimilar ones apart. Hyperparameters were selected through ablation over model depth and dimensionality (Appendix F).

5 Results and Analysis

We follow the official SemEval-2026 protocol.

Task A (Triplets): Primary metric: Accuracy (choose a candidate closer to the anchor). We also examine the confidence between the two similarities to assess the reliability of the model's predictions.

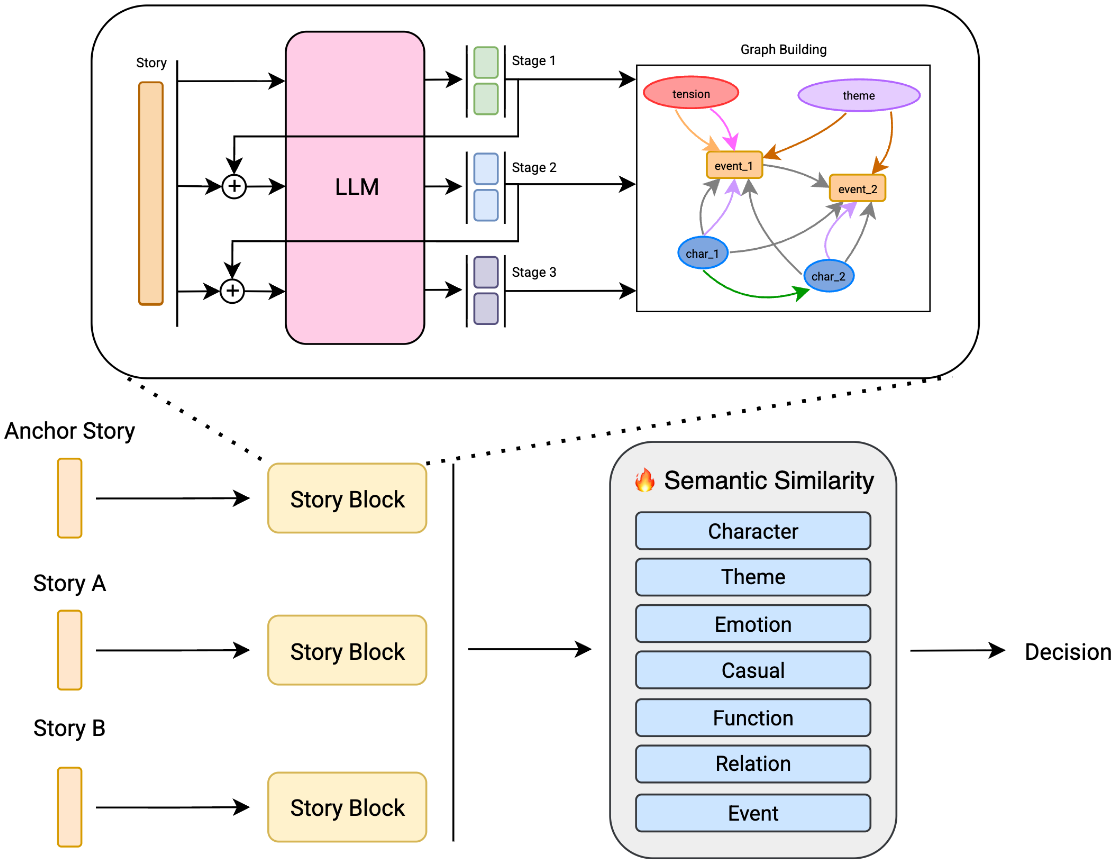


Figure 2: Architecture Diagram for Semantic Similarity Approach.

Table 1: Results on the official test set (Task A).

Model	Acc. (%)	Conf.
LLM Baseline	67.50	-
Semantic Sim.	56.58	0.0533
GAT	57.11	0.2350

Task B (Singles): The organizers perform this using retrieval-based ranking, where cosine similarity between submitted story embeddings is compared against human-annotated narrative similarity judgments.

Baseline. To contextualize the performance of our graph-based similarity measures, we evaluate a large language model (*Gemini 2.5 Flash-Lite*) prompted directly with the anchor, Text A, and Text B to judge which candidate is narratively closer to the anchor. This serves as a reference for how a strong embedding-centric model performs when narrative similarity is inferred implicitly without explicit structural decomposition.

Results on the official test set are reported in Table 1. The LLM baseline achieves the highest accuracy at 67.50%. Among our structured approaches, the GAT model (57.11%) slightly outperforms the semantic model (56.58%), and maintains considerably higher confidence (0.2350 vs. 0.0533), suggesting that learned graph representations produce more decisive similarity margins even when accuracy is comparable. Both structured approaches remain close to chance on the test set despite stronger performance on the training mixture. Our analysis suggests that this gap is not due to the test set being inherently more difficult, but rather due to a mismatch between the synthetic-heavy training distribution and the human-authored narratives present in both development and test data. Ablation studies are provided in Appendix F.

Leaderboard Standing. On the official shared-task leaderboard we placed 43rd out of 47 teams on Track A (triplet comparison) and 27th out of 28 teams on Track B (single-story embeddings).

Discussion. Our models perform well on the training mixture but fail to generalize to the official test set. Rather than indicating that the test

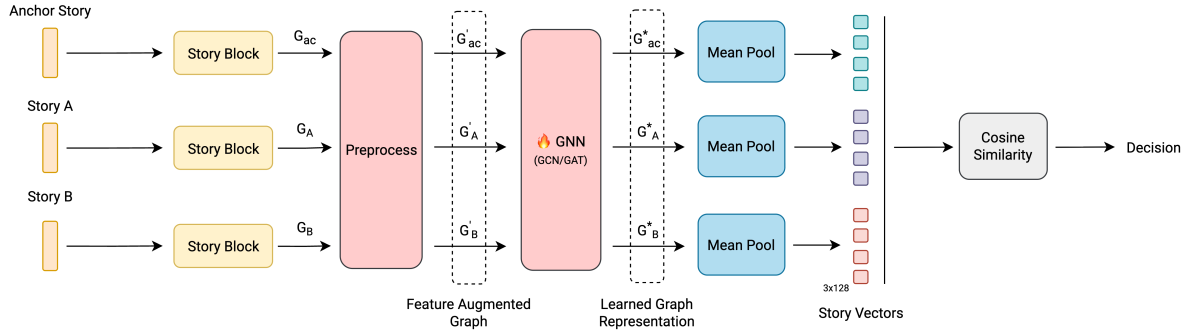


Figure 3: Architecture Diagram for GNN Approach.

set is inherently more challenging, our analysis suggests that the models have primarily learned patterns specific to the dominant synthetic portion of the training data. Since the synthetic distribution differs substantially from the human-authored narratives present in both development and test sets, the learned similarity signals do not transfer effectively. This results in a sharp drop in performance when evaluated on the target distribution.

5.1 Distributional Shift Analysis

To better understand the generalization gap, we analyzed the distribution of narratives across three subsets: (i) the human-authored development subset of 200 examples, (ii) the synthetic subset of 1,897 examples, which constitutes the majority of our training mixture, and (iii) the official test set of 400 examples.

We observe that the development and test distributions are relatively well aligned in embedding space, whereas the synthetic subset exhibits substantial divergence, with a distinct centroid and broader spread. This indicates that the model was primarily exposed to a distribution during training that does not reflect the target evaluation distribution.

As a result, the strong performance observed during training is better explained as fitting to the dominant synthetic-data distribution rather than learning generalizable narrative similarity signals. The subsequent drop in performance on the test set can therefore be attributed to this distributional mismatch, where learned patterns do not transfer to the human-authored narrative structure present in both development and test data.

5.2 Qualitative Analysis

We examine three development-set triplets where each approach succeeds while the others fail.

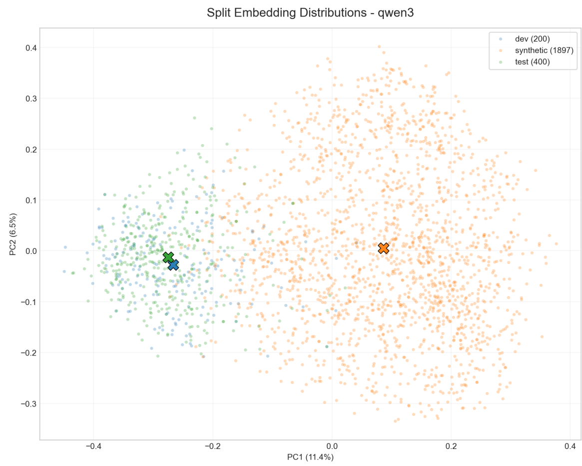


Figure 4: PCA projection of narrative embeddings across splits. Development (blue) and test (green) distributions are closely aligned, while the synthetic training subset (orange) occupies a distinct region, indicating a significant distributional mismatch. Markers denote split centroids.

LLM Baseline Succeeds. The anchor and correct candidate share a high-level narrative arc—disruption, escalating tension, and emotional resolution—but differ in surface entities and event granularity. The LLM captures shared narrative intent holistically, while both structured models are misled: the semantic model by overlapping perfacet scores with the distractor, and the GAT by the correct candidate's sparser graph.

GAT Succeeds. Both candidates share thematic motifs with the anchor but diverge in causal structure. The correct candidate mirrors the anchor's causal chain, while the distractor follows a different progression. The GAT identifies the match by propagating information along causal edges and amplifying structurally salient connections, whereas the LLM and semantic model are misled by surface thematic overlap.

Semantic Model Succeeds. The correct candidate shares strong thematic and emotional parallels with the anchor but has a sparser graph. The semantic model's per-dimension scoring captures these correspondences directly, while the GAT preferentially weights the denser distractor and the LLM is misled by the distractor's richer descriptive language.

6 Conclusion

This work introduces STORYNET, a hybrid framework that combines symbolic narrative decomposition with graph-based representation learning to examine narrative similarity beyond purely embedding-centric approaches. By representing stories as heterogeneous graphs composed of characters, events, and themes, the framework enables structured modeling of relational, causal, and emotional patterns within narratives.

The semantic facet-based model captures explicit thematic, emotional, and causal correspondences and offers interpretable alignment signals through structured scoring. At the same time, its dimension-wise formulation limits its ability to model complex cross-facet interactions and global coherence.

The GAT-based model, in contrast, learns relational dependencies directly from graph structure, allowing it to encode higher-order interactions among narrative elements. This representation can reflect global structural alignment but is sensitive to graph sparsity, density, and connectivity patterns, which can influence stability and generalization.

Taken together, these results illustrate both the potential and the constraints of integrating symbolic narrative structure with learned graph representations. Rather than replacing embedding-based similarity, structured graph modeling offers an alternative analytical lens—one that foregrounds relational organization and interpretable narrative components while introducing new modeling tradeoffs.

References

Apoorv Agarwal, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2013. SINNET: Social interaction network extractor from text. In *Proceedings of IJCNLP 2013: System Demonstrations*, pages 33–36.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Maitreya Prafulla Chitale, Uday Bindal, Rajakrishnan P Rajkumar, and Rahul Mishra. 2025. DiscoGraMS: Enhancing movie screen-play summarization using movie character-aware discourse graph. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 954–965, Albuquerque, New Mexico. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2026 Task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

Rada Mihalcea and Andras Csomai. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 775–780.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Karl Pichotta and Raymond Mooney. 2016. Statistical script learning with recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 11–16, Austin, TX. 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*, pages 3982–3992. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relationships in narrative summaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Lucien Tesnière. 1959. *Eléments de syntaxe structurale*. Klincksieck, Paris.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *Preprint*, arXiv:1710.10903.

Pancheng Wang, Shasha Li, Kunyuan Pang, Liangliang He, Dong Li, Jintao Tang, and Ting Wang. 2022. Multi-document scientific summarization from a knowledge graph-centric view. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 6222–6233, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

A Narrative Facet Taxonomy

This appendix provides formal definitions for all categorical terms used in our narrative facet extraction pipeline across the three stages.

A.1 Stage 1: Character and Relationship Terms

Functional Narrative Roles Characters are assigned one or more functional roles based on their narrative function:

- **Protagonist**: The central character whose journey drives the narrative arc; readers follow their perspective and goals.
- **Antagonist**: A character who opposes the protagonist's objectives, creating conflict and tension within the story.
- **Mentor**: A guiding figure who provides wisdom, training, or resources to help the protagonist on their journey.
- **Victim**: A character who suffers harm or injustice, often motivating action from other characters.
- **Catalyst**: A character whose actions or presence triggers significant plot events or character transformations.
- **Witness**: A character who observes and reports events, often serving as a narrative lens or moral compass.
- **Trickster**: A disruptive character who uses cunning, deception, or humor to challenge the status quo.
- **Helper**: A supporting character who aids the protagonist in achieving their goals.
- **Threshold Guardian**: A character who tests the protagonist at critical junctures, often guarding transitions between story phases.

Agency Levels Agency measures how much a character actively drives plot progression:

- **High**: Character makes consequential decisions that directly shape narrative outcomes.
- **Medium**: Character takes meaningful actions but within constraints set by others or circumstances.
- **Low**: Character primarily reacts to events or is acted upon by others.

Character Types

- **Human**: A human character with typical human capabilities.
- **Animal**: A non-human creature, which may or may not exhibit anthropomorphic traits.
- **Supernatural**: A being with powers or origins beyond natural explanation (e.g., gods, ghosts, magical creatures).
- **Abstract**: A personified concept or force (e.g., Death, Fate, Nature).

Relationship Types

- **Family**: Blood relations or adoptive/chosen family bonds.
- **Romantic**: Relationships involving romantic attraction or partnership.
- **Professional**: Work-based or transactional relationships.

- **Adversarial**: Relationships defined by opposition, conflict, or enmity.
- **Stranger**: Characters who lack prior connection or familiarity.

Relationship Evolution

- **Strengthens**: The bond between characters deepens over the narrative.
- **Weakens**: The relationship deteriorates or becomes more distant.
- **Transforms**: The fundamental nature of the relationship changes (e.g., adversary to ally).
- **Remains Static**: The relationship maintains its initial state throughout.

Power Dynamics

- **Equal**: Neither character holds significant power over the other.
- **First Dominant**: The first character holds greater influence, control, or authority.
- **Second Dominant**: The second character holds greater influence, control, or authority.

A.2 Stage 2: Event and Causality Terms

Event Types

- **Action**: Physical activities or behaviors undertaken by characters.
- **Dialogue**: Verbal exchanges or conversations between characters.
- **Internal**: Mental processes, realizations, or emotional shifts within a character.
- **Environmental**: External occurrences in the setting (e.g., storms, discoveries).
- **Social**: Events involving group dynamics, societal structures, or collective actions.

Narrative Functions Events are categorized by their structural role in the story arc:

- **Inciting Incident**: The event that disrupts the status quo and sets the main conflict in motion.
- **Rising Action**: Events that escalate tension and develop complications.
- **Climax**: The point of highest tension where the central conflict reaches its peak.
- **Falling Action**: Events following the climax that begin resolving tensions.
- **Resolution**: The concluding events that establish a new equilibrium.

Causal Relationship Types

- **Direct Cause**: Event A immediately and necessarily produces Event B.
- **Enabling Condition**: Event A creates circumstances that allow Event B to occur.
- **Catalyst**: Event A accelerates or triggers Event B without being sufficient alone.
- **Consequence**: Event B is a downstream result of Event A, possibly mediated by other factors.

Entity Semantic Fields Objects and entities within events are classified by narrative function:

- **Tool of Agency**: Objects that enable characters to act (e.g., weapons, vehicles).
- **Symbol of Status**: Objects representing power, wealth, or social position.
- **Relationship Marker**: Objects signifying bonds between characters (e.g., rings, letters).
- **Threshold Object**: Objects associated with transitions or boundaries (e.g., keys, maps).
- **Resource**: Objects of material value or utility that characters seek or protect.
- **Obstacle**: Objects that impede character progress or goals.
- **Catalyst Object**: Objects whose introduction triggers significant plot developments.

A.3 Stage 3: Emotional and Thematic Terms

Emotional States Characters' emotional conditions are tracked using the following states: *hopeful, fearful, angry, sad, joyful, conflicted, determined, defeated, surprised, disgusted, contemptuous, ashamed, guilty, proud, envious, grateful, relieved, confused, betrayed, loving, hateful, peaceful, anxious*.

Emotional Change Patterns

- **Gains Hope**: Character transitions toward optimism or positive expectation.
- **Loses Hope**: Character transitions toward pessimism or despair.
- **Maintains State**: Character's emotional state remains consistent.
- **Transforms**: Character undergoes fundamental emotional metamorphosis.

Thematic Abstraction Levels

- **Universal**: Themes applicable across cultures and time periods (e.g., love, death, justice).
- **Cultural**: Themes tied to specific cultural contexts or value systems.
- **Situational**: Themes arising from particular circumstances within the narrative.

Resolution Types

- **Victory**: The protagonist achieves their primary goal.
- **Defeat**: The protagonist fails to achieve their primary goal.
- **Compromise**: Partial fulfillment where goals are modified or shared.
- **Transformation**: Resolution through fundamental change in the protagonist or situation.
- **Ambiguous**: The outcome remains unclear or open to interpretation.
- **Tragic**: Resolution involving significant loss despite potential moral victory.

Emotional Impact The overall emotional effect on the audience:

- **Cathartic**: Provides emotional release or purification.
- **Disturbing**: Leaves the audience unsettled or troubled.
- **Uplifting**: Inspires positive feelings or hope.
- **Melancholic**: Evokes sadness or wistful reflection.
- **Ambivalent**: Produces mixed or conflicting emotional responses.

B Narrative Graph Construction Details

This appendix provides the full specification of graph nodes and edges summarized in the main text.

Graph Nodes. Three major types of nodes are defined:

- **Characters:** Each character extracted in Stage 1 is mapped to a graph node, enriched with attributes such as functional role, agency level, and key descriptors.
- **Events:** Events from Stage 2 are represented as nodes capturing their description, type (e.g., action, dialogue, internal), emotional tone, and narrative function.
- **Themes:** High-level thematic elements identified in Stage 3 are incorporated as abstract nodes, allowing the graph to encode both concrete and conceptual dimensions of the narrative.

Graph Edges. Edges capture different types of relationships:

- **Character–Character:** Edges encode social ties (*family, romantic, adversarial*), including attributes for relationship evolution and power dynamics.
- **Character–Event:** Characters are connected to events they participate in, with directionality and roles (e.g., *primary actor*, *supporting participant*).
- **Event–Event:** Causal chains from Stage 2 are translated into directed edges, annotated with type (*direct cause, enabling condition, consequence*) and strength.
- **Theme–Event:** Each theme node is connected to all event nodes in the graph. Since thematic elements are identified by analyzing the narrative as a whole rather than being localized to specific events, this global connectivity reflects the pervasive nature of themes across the story's event structure.

C Narrative Facet Extraction: Stage Details

This appendix provides the full descriptions of each extraction stage summarized in the main text.

Stage 1: Character and Relationship Extraction. The first stage identifies the set of characters and their functional narrative roles (*protagonist, antagonist, mentor, victim*, etc.). Each character is assigned a unique identifier (char_1, char_2, ...), agency levels (*high/medium/low*), and key descriptive attributes. In addition, pairwise relationships are extracted, annotated with type (*family, romantic, adversarial, professional*), evolution (*strengthens, weakens, transforms*), and power dynamics (*equal, first_dominant, second_dominant*). These details serve as the foundational layer for constructing story graphs.

Stage 2: Event and Causality Extraction. The second stage focuses on the dynamic structure of the narrative. Events are extracted with attributes such as description, participants, emotional tone, narrative function (*inciting incident, climax, resolution*), and temporal markers. Entities involved in events are semantically categorized (*tool_of_agency, resource, obstacle*). The system then infers causal chains between events, including relationship type (*direct cause, enabling condition, consequence*) and temporal gap. Parallel narrative threads are identified where applicable, with explicit convergence points.

Stage 3: Emotional and Thematic Analysis. The third stage captures higher-level abstractions. Emotional trajectories track the evolving states of characters across events (*hopeful, fearful, conflicted, joyful, defeated*, etc.) along with the intensity and direction of emotional change. Thematic elements are identified at different levels of abstraction (*universal, cultural, situational*), noting how themes manifest in the narrative and how they are resolved (*victory, defeat, compromise, tragic destruction*). Finally, a tension–resolution pattern is constructed, specifying the central tension, escalation events, crisis point, resolution type, and emotional impact (*cathartic, disturbing, uplifting, melancholic*).

D Extraction Prompts

Below are the prompts used for each stage of narrative facet extraction. Each prompt is formatted for instruction-tuned LLMs and constrains outputs to structured JSON schemas.

D.1 Stage 1: Character and Relationship Extraction

```
System: You are a helpful assistant. You must
output ONLY VALID JSON that follows the exact
schema provided. Do not include explanations,
notes, or extra text outside of the JSON object.

User: Analyze the following story and extract
character information in the specified JSON format.

Story: {story_text}

Extract the following information:
1. All characters mentioned in the story
2. Their functional narrative roles
3. Key relationships between characters
4. Character agency levels (high/medium/low)

Return ONLY valid JSON in this exact format:
{
  "characters": [
    {
      "id": "char_x",
      "name": "Character Name or Description",
    "functional_roles": ["protagonist", "victim"],
      "agency_level": "high|medium|low",
      "key_attributes": ["brave", "conflicted"],
      "character_type": "human|animal|...",
      "aliases": ["nickname", "title"]
    }
  ],
  "relationships": [
    {
      "first": "char_1",
      "second": "char_2",
     "relationship_type": "family|romantic|...",
    "relationship_evolution": "strengthens|...",
    "power_dynamic": "equal|first_dominant|...",
      "evidence_text": "quote from story"
    }
  ]
}

Functional roles options: protagonist, antagonist,
catalyst, victim, mentor, witness, trickster,
helper, threshold_guardian
```

D.2 Stage 2: Event and Causality Extraction

```
System: You are a precise narrative analyst. You
must output ONLY valid JSON. Focus ONLY on major
plot-driving events. Do NOT list minor actions.

User: Analyze the story for events and their
causal relationships. Use the character IDs below.

Story: {story_text}
Character IDs: {character_ids}

Extract:
1. Key structural events (inciting incident,
   major action, climax, resolution)
2. Causal relationships between events
3. Any parallel plot threads

Return ONLY valid JSON in this exact format:
{
  "events": [
    {
      "id": "event_1",
      "description": "Brief description",
    "event_type": "action|dialogue|internal|...",
      "participants": ["char_1", "char_2"],
```

```
        "primary_actor": "char_1",
        "emotional_tone": "positive|negative|...",
      "narrative_function": "inciting_incident|...",
        "time_marker": "morning|evening|...",
        "entities_involved": [
          {
            "entity": "gun",
            "semantic_field": "tool_of_agency",
         "narrative_significance": "high|medium|low"
          }
        ]
      }
    ],
    "causal_chains": [
      {
        "cause_event": "event_1",
        "effect_event": "event_2",
        "relationship_type": "direct_cause|...",
        "strength": "strong|medium|weak",
        "temporal_gap": "immediate|short|long"
      }
    ],
    "parallel_threads": [
      {
        "thread_id": "thread_1",
        "events": [event IDs],
        "convergence_point": "event_x"
      }
    ]
}
```

D.3 Stage 3: Emotional and Thematic Analysis

```
System: You are a helpful assistant. You must
output ONLY valid JSON that follows the exact
schema provided.

User: Analyze the emotional trajectories and
thematic elements of the story.

Story: {story_text}
Characters: {characters}
Events: {events}

Extract emotional and thematic information:
{
  "emotional_trajectories": [
    {
      "character_id": "char_1",
      "trajectory": [
        {
          "event_id": "event_1",
          "emotional_state": "hopeful",
          "intensity": "high|medium|low",
          "emotional_change": "gains_hope|..."
        }
      ]
    }
  ],
  "thematic_elements": [
    {
      "theme": "love_vs_duty",
    "abstraction_level": "universal|cultural|...",
      "manifestation": "How theme is expressed",
      "resolution": "duty_wins|love_wins|..."
    }
  ],
  "tension_resolution_pattern": {
    "central_tension": "Main opposing forces",
    "tension_escalation": ["event_2", "event_5"],
    "crisis_point": "event_9",
    "resolution_type": "victory|defeat|...",
   "emotional_impact": "cathartic|disturbing|..."
  }
}

Emotional states: hopeful, fearful, angry, sad,
joyful, conflicted, determined, defeated,
surprised, disgusted, contemptuous, ashamed,
guilty, proud, envious, grateful, relieved,
confused, betrayed, loving, hateful, peaceful,
anxious
```

E GNN Encoder Variants

This appendix details the specific architectural configurations of the two encoder variants evaluated: GCN and GAT. Both models take 768-dimensional node features as input and produce a 1024-dimensional graph-level embedding.

E.1 Graph Convolutional Network (GCN)

We implement a two-layer GCN with ReLU activations and dropout applied between layers (see Figure 5a):

- **Layer 1:** $768 \rightarrow 2056$, ReLU
- **Layer 2:** $2056 \rightarrow 1028$, ReLU

A global mean pooling operation aggregates node embeddings into a graph-level representation. This pooled vector is then linearly projected $1028 \rightarrow 1024$ to obtain the final graph embedding.

E.2 Graph Attention Network (GAT)

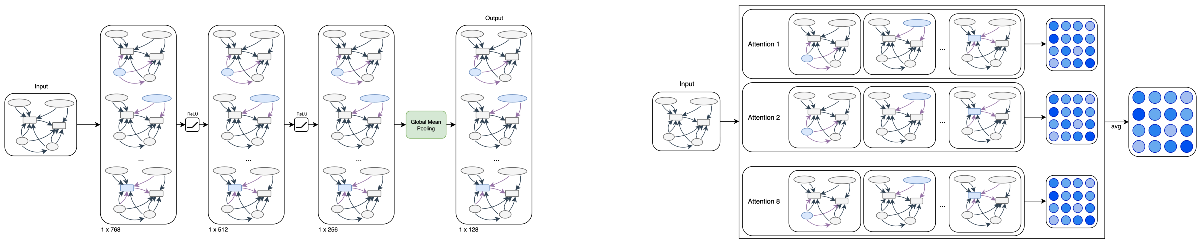
We implement a two-layer multi-head GAT with ELU activations and dropout (see Figure 5b):

- **Layer 1:** 8 attention heads, each of size 512. Concatenated output dimension: $512 \times 8 = 4096$. Mapping: $768 \rightarrow 4096$, ELU
- **Layer 2:** $4096 \rightarrow 4096$, ELU

The outputs of the 8 attention heads in the first layer are concatenated.After message passing, global mean pooling aggregates node embeddings, followed by a linear projection $4096 \rightarrow 1024$ to produce the final graph-level embedding.

F Ablations

All ablation results reported in this section are evaluated on an in-house 80–10–10 train–validation–test split of the compiled development set, not on the official test data. We did not ablate on the official test set due to the limited number of submissions permitted by the task organizers.



(a) Graph Convolutional Network (GCN) variant.

(b) Graph Attention Network (GAT) variant.

Figure 5: GNN Encoder Variants

F.1 Loss functions.

We compared InfoNCE and triplet margin loss under an (anchor, positive, negative) training setup. In our experiments, InfoNCE yielded weaker performance. One possible explanation is that, in this task, negatives are not uniformly hard nor sampled in large batches. Each triplet contains a single candidate negative whose similarity to the anchor varies substantially across samples. Under these conditions, the temperature-scaled contrastive objective may receive uneven gradient signals, potentially limiting consistent optimization of the anchor–positive alignment.

In contrast, the triplet margin loss explicitly optimizes the relative ordering between anchor, positive, and negative examples. This formulation may be better suited to pairwise comparative settings where negative difficulty varies across instances. By enforcing a fixed separation margin, it appears to provide more stable optimization dynamics in our setup, converging faster and yielding higher accuracy and confidence scores (Table 2).

F.2 GNN variant

We also treat the choice of message-passing architecture as an ablation. Using the same 1024-dimensional graph-level embedding and triplet loss, the attention-based GAT substantially outperforms the GCN (Table 1). This suggests that the ability to assign differentiated weights to neighboring nodes may be particularly important in heterogeneous story graphs. In such graphs, it is plausible that only a subset of nodes—such as those associated with central tension or major turning points—carry the most diagnostic signal for narrative similarity. If so, mechanisms that allow selective emphasis over relational structure could be advantageous in capturing these dynamics.

Table 2: Ablation: Loss functions on GAT for 75 epochs

Objective	**Acc. (%)**	**Conf.**
InfoNCE	85.6	0.1630
Triplet Margin	94.21	0.2835

Table 3: Ablation: GNN hidden dims.

GCN Layout	**Acc. (%)**	**Conf.**
768→256→512→1024	70.44	0.1379
768→1024→512→1024	80.01	0.1427
768→2056→1028→1024	94.21	0.2835

F.3 Hidden dimensions (GNN)

We observed that non-monotonic "bottleneck-then-expansion" configurations and overly wide early layers yielded weaker performance compared to a pyramidal taper. One possible interpretation is that an early bottleneck may discard useful variance before sufficient graph context is aggregated, and subsequent dimensional expansion may increase parameter count without recovering lost information, potentially encouraging overfitting.

In contrast, a steady dimensional decrease may preserve salient features during initial message passing while gradually compressing representations in alignment with the final pooling stage. This configuration appears to regularize capacity more effectively and was associated, in our experiments, with more stable optimization and improved accuracy. (Table 3).