MarSan at SemEval-2026 Task 4: Narrative Similarity via Sentence-BERT Metric Learning with Triple-Derived Losses

Maryam Najafi
Department of Computer Science and Information Systems
University of Limerick
Ireland
najafi.maryam@ul.ie

Ehsan Tavan
NLP Department,
Part AI Research Center
Tehran
Iran
ehsan.tavan@partdp.ai

Simon Colreavy
Department of Computer Science and Information Systems
University of Limerick
Ireland
simon.colreavy@ul.ie

Abstract

We describe our research to SemEval-2026 Task 4 on Narrative Story Similarity and Narrative Representation Learning (NSNRL). The shared task defines narrative similarity through comparative judgments over triples consisting of an anchor story and two candidates, where systems determine which candidate is narratively closer (Track A), and must output story embeddings whose cosine distances reproduce the same ordering under withheld evaluation triples (Track B). We implement a unified representation-learning approach based on a Sentence-BERT bi-encoder trained with triple-derived metric learning objectives, combining in-batch contrastive learning with explicit triplet and margin-ranking constraints. Track A is solved by direct cosine comparison between the anchor embedding and each candidate embedding, while Track B outputs normalized story vectors from the same encoder without any additional test-time modelling. During evaluation, we achieve 65.00% accuracy on Track A and 65.50% on Track B. These results suggest that a single, well-aligned bi-encoder can perform competitively across both tracks while remaining computationally efficient.

1 Introduction

Narrative similarity concerns whether two stories correspond at an abstract level—through shared themes, causal organisation, character goals, and outcomes—rather than via surface lexical overlap (Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016; Ammanabrolu et al., 2020). A recurring insight across computational narratology and NLP is that such resemblance is often grounded in *structure*: stories can be compared through aligned event sequences, role configurations, and causal progressions (Reiter, 2014; Reiter et al., 2014). Accordingly, narrative similarity has been theorized as the existence of an appropriate *common summary* that preserves what is essential and shared while abstracting away incidental detail (Kypridemou and Michael, 2013, 2014). Empirical studies also show that human similarity judgments are both tractable and influenced by multiple factors, and that they can be collected at scale through annotation and crowdsourcing, with annotators often explaining their choices by pointing to similarities in plot and outcomes (Nguyen et al., 2014). In applied settings, these abstractions manifest in naturally occurring retellings—such as movie remakes—where narratives remain similar despite changes in style, entities, and presentation (Chaturvedi et al., 2018).

SemEval-2026 Task 4 embraces this view by operationalizing narrative similarity as a *comparative* judgment problem (Hatzel et al., 2026). Each instance provides an anchor story and two candidates, and systems must predict which candidate is narratively closer to the anchor (Track A). Track B evaluates the learning of narrative representation by testing whether cosine distances between single-story embeddings reproduce the same triple preference relations. This pairwise framing mitigates the calibration difficulties of absolute similarity scoring and connects directly to classical comparative judgment and modern learning-to-rank formulations (Thurstone, 1927; Bradley and Terry, 1952; Joachims, 2002).

MarSan participates in both tracks under a unified modelling principle: if narrative similarity is observed through relative comparisons, training should directly impose *relative constraints* on an embedding space evaluated with the same cosine similarity used at inference time (Weinberger and Saul, 2009; Schroff et al., 2015). For Encoder we used Alibaba's *gte-large-en-v1.5* which is an English embedding model supporting long contexts up to 8192 tokens via a transformer++ backbone (BERT + RoPE + GLU). It reports strong Massive Text Embedding Benchmark (MTEB) performance within its size class. The model is trained with a multi-stage pipeline that progressively ex-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 228–234

tends MLM pretraining to longer lengths before contrastive pretraining/fine-tuning (Zhang et al., 2024; Li et al., 2023). We therefore adopt a Sentence-BERT bi-encoder architecture (Reimers and Gurevych, 2019) to this model to encode each story summary into a dense, L2-normalized vector. We train with a composite metric-learning objective: in-batch contrastive learning (denoted $\mathcal{L}_{\text{MN}}$) shapes the global geometry required for Track B generalization, while an explicit local ordering constraint (denoted $\mathcal{L}_{\text{tri}}$) enforces the anchor–positive vs. anchor–negative preference required for Track A decisions (van den Oord et al., 2018; Chen et al., 2020; Schroff et al., 2015). To broaden coverage of narrative phenomena, we mix human-annotated Wikipedia plot summaries with LLM-generated synthetic triples. On the official evaluation, MarSan achieved 65.00% accuracy on Track A and 65.50% on Track B, suggesting that a single, well-aligned bi-encoder can remain competitive across comparative decision-making and representation learning while maintaining practical efficiency. Our code is available online at: `https://github.com/MaryNJ1995/narrative_similarity2026`.

2 Background

While semantic similarity often tracks topical overlap or paraphrase equivalence, narrative similarity targets higher-level correspondences such as event structure, causal dependencies, character intentions, and resolutions (Chambers and Jurafsky, 2008; Mostafazadeh et al., 2016). From a modelling standpoint, this shift motivates representational choices that are sensitive to *what happens* and *why it happens*, not only to shared vocabulary. Work in computational narratology has formalized these correspondences through structural alignment, where stories are compared by aligning event sequences and narrative relations across documents (Reiter, 2014; Reiter et al., 2014). A complementary behavioural and computational line of work treats narrative similarity as the extent to which two stories admit a *common summary*—a single account that captures their shared gist (Kypridemou and Michael, 2013, 2014). Most studies indicate that both expert and non-expert annotators can produce reliable similarity judgments; however, these judgments are typically grounded in abstract narrative properties—such as a shared conflict and outcome—rather than surface-level overlap (Nguyen et al., 2014). Comparable correspondences also arise outside controlled settings, for instance in movie remakes that retain a common plot backbone while changing style, characters, and setting (Chaturvedi et al., 2018). Recent NLP research has further extended similarity modelling to personal and spoken narratives, underscoring the importance of capturing narrative-level structure that goes beyond lexical or purely topical similarity (Saldias and Roy, 2020; Shen et al., 2023).

NSNRL operationalizes narrative similarity via triple-wise comparisons rather than absolute similarity scores. Comparative evaluation is well-motivated when absolute scaling is difficult: classical models of comparative judgment (e.g., Thurstone and Bradley–Terry) and learning-to-rank formulations provide principled tools for preference learning and evaluation (Thurstone, 1927; Bradley and Terry, 1952; Joachims, 2002). For NSNRL, bi-encoders are appealing because they eliminate the need for joint cross-attention during inference and because Track B explicitly rewards coherent global geometry under cosine similarity. Since the benchmark measures relative preferences, metric learning naturally offers a way to construct an embedding space that adheres to ordering constraints (Weinberger and Saul, 2009; Schroff et al., 2015; van den Oord et al., 2018; Chen et al., 2020).

3 Task and Data

SemEval-2026 Task 4 evaluates how closely two stories match in abstract causal progression and plot development while discounting surface details such as names, settings, and objects (Hatzel et al., 2026).

3.1 Official Dataset

NSNRL comprises two tracks. In Track A, each instance is a triple (a, c_1, c_2) with an anchor story a and two candidates; systems predict which candidate is narratively closer. The official Track A dataset consists of 39 instances in the sample split, 200 in the development split, and 400 in the test split . In Track B, systems output one embedding per story and are evaluated by whether cosine distance preserves the same ordering, i.e., $\text{dist}(a, c^+) < \text{dist}(a, c^-)$. For Track B test, the organizers withhold the evaluated triple groupings to prevent test-time tuning. Track B contains 1,436 unique stories in total (108 sample, 479 dev, 849 test).

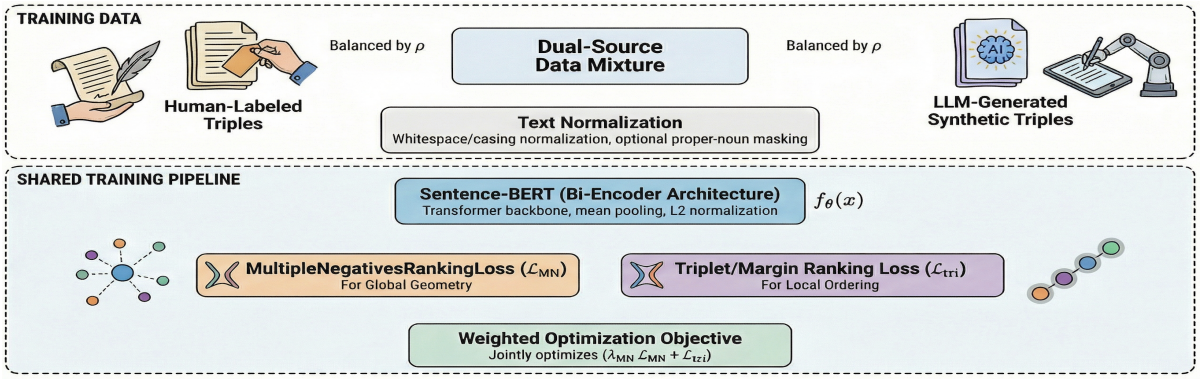

Figure 1: Overview of MarSan's unified training pipeline for SemEval-2026 Task 4.

The stories are sourced from the English portion of TELL-ME-AGAIN and filtered to 4-8 sentence summaries (Hatzel and Biemann, 2024). Since no large official training split is released, the organizers provide synthetic training data produced by prompting multiple LLMs to generate Wikipedia-style plots and paired variants: a narratively similar rewrite that mirrors the plot structure while changing surface details, and a narratively distant story with different conflict and resolution. The synthetic release contains 1,900 triples.

3.2 Synthetic Data Construction and Integration

Given the absence of a large official training split, we train primarily with LLM-generated synthetic supervision and use the official development set only for model selection. Our synthetic resources are designed to mirror the Track A comparison setting and to be directly reusable for metric learning with a bi-encoder.

We use two complementary synthetic datasets. The first is a Track A-style triple dataset containing an `anchor_text` and two candidates (`text_a`, `text_b`) with a deterministic label (`text_a_is_closer`). The second is a contrastive dataset containing an anchor story together with a narrative-preserving rewrite (positive) and a narratively distant variant (negative). From these formats we construct anchor–positive pairs (a, p) for in-batch contrastive learning and triplets (a, p, n) for an explicit margin-based ordering constraint, aligning training with the triple-wise evaluation used in both tracks.

To reduce reliance on lexical shortcuts, the synthetic generator produces negatives with graded difficulty. Easy negatives differ strongly in topic or setting; medium negatives retain a loose thematic connection while diverging in plot mechanics; and hard negatives preserve high surface overlap but alter narrative causality or outcomes. In our local setup, the training pool contains 1,700 synthetic instances, while the official development split contains 200 labeled triples and is used exclusively for early stopping and checkpoint selection.

4 System Overview

Our system is a bi-encoder that maps each story summary x to a dense vector $f_\theta(x) \in \mathbb{R}^d$ and uses cosine similarity

$$\mathrm{sim}(x, y) = \frac{f_\theta(x)^\top f_\theta(y)}{\|f_\theta(x)\|\|f_\theta(y)\|} \quad (1)$$

to compare stories. The same encoder is used for both tracks. Track A predicts the closer candidate by comparing $\mathrm{sim}(a, c_1)$ and $\mathrm{sim}(a, c_2)$, while Track B exports the final normalized embedding $z = \frac{f_\theta(x)}{\|f_\theta(x)\|}$ for each story.

4.1 Sentence-BERT Bi-Encoder Formulation

We use a transformer encoder as the token-level backbone and convert the variable-length sequence representation into a fixed-size sentence embedding. In the Sentence-BERT formulation, we encode each story independently and compute similarity using cosine similarity between embeddings. This design scales efficiently because it avoids cross-attention between the anchor and candidates , and it matches Track B evaluation, which is based on cosine distances between story embeddings.

Concretely, for a story x tokenized into T tokens, the backbone produces contextual token vectors $h_1, \ldots, h_T$. We compute a pooled representation

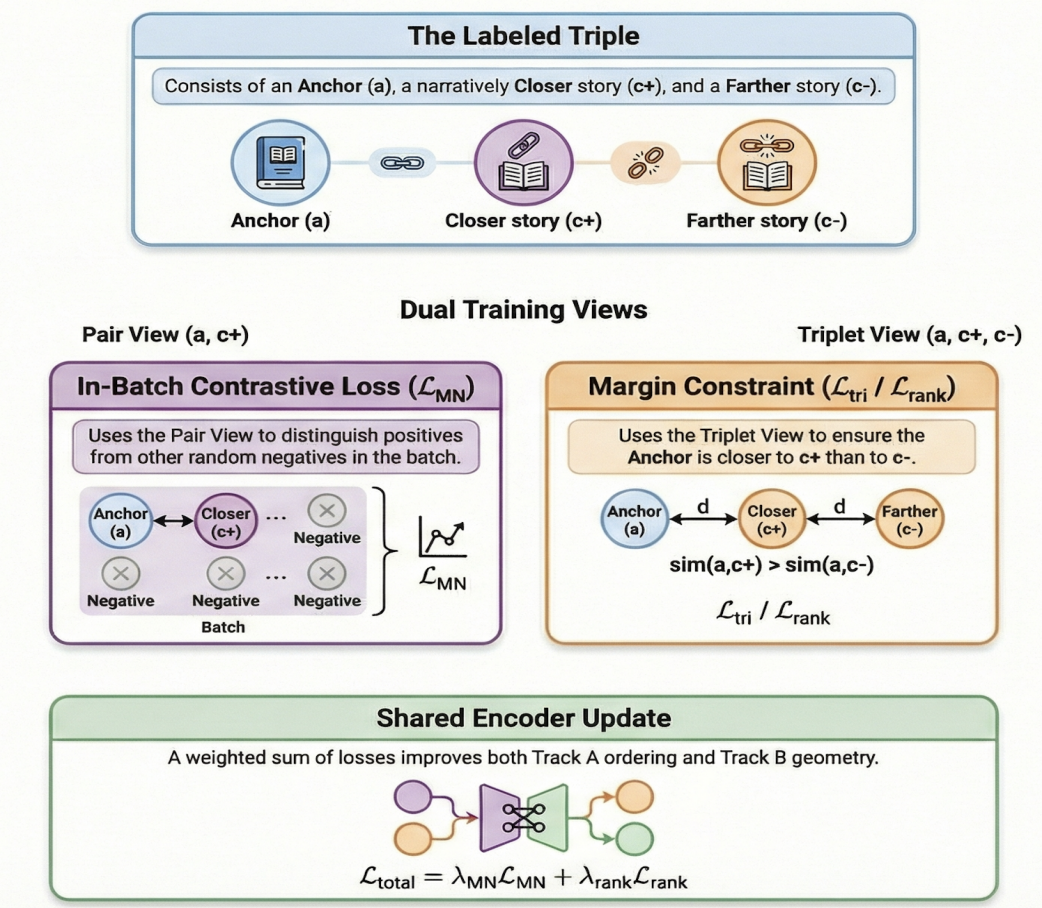

Figure 2: Loss overview: in-batch contrastive shaping plus explicit triple-wise ordering constraints.

using mean pooling over the non-padding tokens,

$$f_\theta(x) = \frac{1}{T'} \sum_{t=1}^{T'} h_t, \tag{2}$$

and apply L2 normalization before cosine similarity computations.

4.2 Loss Design: Aligning Training with Triple-Based Evaluation

For each labeled triple we denote the anchor by a, the narratively closer candidate by c^+, and the narratively farther candidate by c^-. We train with a composite objective that combines in-batch contrastive learning with explicit triple constraints.

In-batch contrastive learning with MultipleNegativesRankingLoss. For a batch of N anchor-positive pairs $\{(a_i, p_i)\}_{i=1}^N$, the loss is

$$\mathcal{L}_{\text{MN}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(a_i, p_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(a_i, p_j)/\tau)}, \tag{3}$$

where τ is a temperature parameter.

Explicit triplet constraint with TripletLoss. We additionally enforce a triplet constraint:

$$\mathcal{L}_{\text{tri}} = \max\left(0, m + d(a, c^+) - d(a, c^-)\right), \tag{4}$$

where m is a margin and $d(\cdot, \cdot) = 1 - \text{sim}(\cdot, \cdot)$ is cosine distance.

Preference-style margin ranking as the Track A surrogate. We can also express the same constraint as a hinge ranking loss:

$$\mathcal{L}_{\text{rank}} = \max\left(0, m - \text{sim}(a, c^+) + \text{sim}(a, c^-)\right). \tag{5}$$

We use $\mathcal{L}_{\text{rank}}$ as an equivalent hinge form of the same local ordering constraint; in our experiments we optimize only one of $\mathcal{L}_{\text{tri}}$ or $\mathcal{L}_{\text{rank}}$ to avoid redundant gradients. Overall objective is as follow:

$$\mathcal{L} = \lambda_{\text{MN}}\mathcal{L}_{\text{MN}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}}. \tag{6}$$

For the final submission, we fine-tune the encoder for 4 epochs with maximum sequence length 256, learning rate 2×10^{-5}, warmup ratio 0.1, weight

Variant	Backbone(s) and role	Track A (%)	Track B (%)
MarSan submission	`Alibaba-NLP/gte-large-en-v1.5` fine-tuned as Sentence-BERT bi-encoder; cosine decision for Track A; L2-normalized embeddings for Track B	**65.00**	**65.50**
Sentence-BERT baseline	`all-MiniLM-L6-v2` as bi-encoder baseline (reference point)	56.25	63.25
Bi-encoder (gte) + $\mathcal{L}_{\text{MN}}$ only (dev)	`gte-large-en-v1.5` bi-encoder trained with MultipleNegativesRankingLoss; cosine decision / normalized embeddings	64.20	65.20
Bi-encoder (gte) + $\mathcal{L}_{\text{tri}}$ only (dev)	`gte-large-en-v1.5` bi-encoder trained with Triplet/MarginRanking only; cosine decision / normalized embeddings	63.60	64.60
Bi-encoder (gte) + dual-source triples (dev)	`gte-large-en-v1.5` bi-encoder trained on human+synthetic mixture (balanced by ρ), composite objective	64.80	65.10
Bi-encoder + teacher re-ranking (explored; dev)	`gte-large-en-v1.5` bi-encoder with optional re-ranking / distillation signals from `ms-marco-MiniLM-L-6-v2`	64.60	65.00
Hard-negative mining with cross-encoder (explored; dev)	Triples augmented with hard negatives selected by `ms-marco-MiniLM-L-6-v1`; bi-encoder trained with composite objective	64.90	65.30

Table 1: Internal model comparison for MarSan: baseline bi-encoder, single-loss ablations, and data/negative-mining augmentations. All variants use cosine similarity for Track A and L2-normalized embeddings for Track B; scores are reported as accuracy (%). ρ also controls the mixing ratio between human-labeled and synthetic training triples.

decay 0.0, and gradient accumulation of 1. We use batch size 16 for anchor–positive contrastive pairs and 8 for triplets. The temperature τ, margin m, and loss weights λ_{MN} and λ_{tri} are defined outside the shown training arguments and are therefore reported only where explicitly specified in the released configuration. Checkpoints are selected on the 200-example development split using Track A accuracy, and the selected checkpoint is then reused unchanged to export Track B embeddings.

5 Results

We initially used Alibaba's model as the primary bi-encoder for both tracks and trained it using triple-derived metric-learning losses. Second, we used a lightweight cross-encoder (`cross-encoder/ms-marco-MiniLM-L-6-v2`) as a teacher and as an optional re-ranker for Track A-style decisions. Table 1 lists the model variants explicitly referenced in our development trace, along with the results reported in the shared-task report and those from our final submissions. For reproducibility, the main detail is that the final MarSan submission relied on Alibaba's Sentence-BERT bi-encoder model, trained with a composite metric-learning objective. Track A predictions were generated via direct cosine similarity between the anchor and the candidate sentences, whereas Track B outputs consisted of L2-normalized embedding vectors.

Table 2 summarizes our incremental experiments as a controlled build-up toward the final MarSan submission. We start from a lightweight SBERT baseline and then strengthen the encoder with Alibaba's large model, which yields the largest single improvement by increasing representation capacity while retaining efficient cosine-based inference. Next, we expand supervision using a dual-source mixture of human-labeled triples and LLM-generated synthetic triples, improving both tracks by exposing the model to a broader set of narrative phenomena. We then add metric-learning losses that align directly with NSNRL evaluation: $\mathcal{L}_{\text{MN}}$ (in-batch contrastive learning) improves the global geometry of the embedding space and benefits Track B, while incorporating $\mathcal{L}_{\text{tri}}$ (triplet/margin ranking) explicitly enforces the local ordering needed for Track A decisions. Finally, text normalization and optional proper-noun masking provide a small but consistent gain, yielding our best configuration at 65.00% on Track A and 65.50% on Track B. Compared to the provided baselines, our unified bi-encoder approach substantially exceeds random chance and the lexical baseline, remains competitive with stronger non-embedding approaches on Track A, and improves upon the story-embedding baseline on Track B. Also an important observation from Table 1 is that

Model / Incremental Change	Track A Acc. (%)	Track B Acc. (%)
Random baseline (coin flip)	50.00	50.00
Story-Embedding baseline (organizers)	–	63.25
GPT-4o-mini prompting baseline (organizers)	67.00	–
Best submission (leaderboard)	**78.00** (COGNAC)	**72.00** (COGNAC)
(1) SBERT bi-encoder (all-MiniLM-L6-v2) baseline	56.25	63.25
(2) + Stronger backbone: gte-large-en-v1.5	62.00	64.50
(3) + Dual-source data (human + synthetic triples), balanced by ρ	63.50	64.90
(4) + Composite objective: $\mathcal{L}_{\mathrm{MN}}$ (in-batch contrastive)	64.20	65.20
(5) + Add local ordering: $\mathcal{L}_{\mathrm{tri}}$ (triplet/margin ranking)	64.80	65.40
(6) + Text normalization & optional proper-noun masking	**65.00**	**65.50**

Table 2: Incremental results for MarSan. Rows (1)–(6) form an ablation-style build-up: each step adds a single component to the previous configuration. The final configuration (6) is our submitted system and achieves the best performance among our bi-encoder variants. We also report key reference points: organizer baselines and the current best leaderboard submissions for each track.

$\mathcal{L}_{\mathrm{MN}}$ improves performance on both tracks more strongly than $\mathcal{L}_{\mathrm{tri}}$ when used alone, suggesting that global embedding geometry is a major driver of narrative similarity performance. At the same time, the best results are obtained by combining both objectives, indicating that global structure and local ranking constraints are complementary rather than strictly task-specific.

6 Discussion and Error Analysis

The comparison between tracks highlights an important methodological distinction: bi-encoders are efficient and can learn meaningful geometry, but they may struggle with fine-grained distinctions that can be resolved by jointly conditioning on both texts. Track A can sometimes be improved with cross-encoder re-ranking because the model can attend to detailed correspondences between an anchor and a candidate, including subtle differences in causality or outcomes. Track B, however, requires a single embedding per story, so improvements must come from better global geometry rather than pair-specific modelling. Errors frequently arise in cases where candidates trade off different similarity components. A candidate may match the anchor's course of action but differ in outcomes, while another may share abstract theme but follow a different event structure. Since the annotation guidelines do not prescribe a fixed weighting among narrative components, a single-vector representation can conflate these dimensions. Another issue would be information sparsity in summaries, where missing plot steps force the model to rely on genre cues or entity-level overlap.

Our approach uses a single-vector embedding per story, which can conflate multiple narrative facets (e.g., outcome similarity vs. event-structure similarity) and may underperform on borderline cases that require fine-grained relational reasoning. In synthetic augmentation, if generated triples violate the intended preference ordering, it can introduce noise, and our filtering heuristics cannot guarantee perfect correctness. Finally, Track A may benefit from cross-encoder re-ranking, but such models increase inference cost and do not transfer to Track B, which requires story-level embeddings.

7 Conclusion

We presented a unified Sentence-BERT bi-encoder for SemEval-2026 Task 4 that solves both comparative narrative similarity (Track A) and narrative representation learning (Track B) challenge within a single cosine-evaluated embedding space. On the official evaluation set, our single-model approach achieves 65.00% accuracy on Track A and 65.50% on Track B, demonstrating that a compact, well-aligned bi-encoder can remain competitive across both tracks while preserving the efficiency and deployability of embedding-based inference. Future work will focus on more faithful hard-negative generation and filtering, multi-facet representations that separate narrative components (e.g., event structure vs. outcome), and distillation from stronger cross-encoders into the bi-encoder to improve fine-grained decisions without sacrificing Track B compatibility.

Acknowledgments

Publication stems from research supported financially by Taighde Éireann – Research Ireland, under Grant Number 18/CRT/6223.

References

Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020. Story realization: Expanding plot events into sentences. In *Proceedings of AAAI*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3–4):324–345.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have I heard this story before? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 673–678. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*.

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.

Hans Ole Hatzel and Chris Biemann. 2024. Tell me again! a large-scale dataset of multiple summaries for the same story. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15732–15741, Torino, Italia. ELRA and ICCL.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of KDD*.

Elektra Kypridemou and Loizos Michael. 2013. Narrative similarity as common summary. In *2013 Workshop on Computational Models of Narrative*, volume 32 of *OpenAccess Series in Informatics (OASIcs)*, pages 129–146. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Elektra Kypridemou and Loizos Michael. 2014. Narrative similarity as common summary: Evaluation of behavioral and computational aspects. *Literary and Linguistic Computing*, 29(4):532–560.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*.

Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2014. Using crowdsourcing to investigate perception of narrative similarity. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM '14)*, pages 321–330. ACM.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*.

Nils Reiter. 2014. *Discovering Structural Similarities in Narrative Texts Using Event Alignment Algorithms*. Phd thesis, Heidelberg University.

Nils Reiter, Anette Frank, and Oliver Hellwig. 2014. An nlp-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605.

Belen Saldias and Deb Roy. 2020. Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 78–86, Online. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of CVPR*.

Jocelyn Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Park, and Cynthia Breazeal. 2023. Modeling empathic similarity in personal narratives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6252, Singapore. Association for Computational Linguistics.

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*.

Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. In *Proceedings of NeurIPS*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.