Compare Results

Old File:		New File:
2024.emnlp-main.339.pdf	versus	**2024_emnlp-main_339.pdf**
13 pages (229 KB)		**14 pages (304 KB)**
10/31/2024 10:32:46 PM		2/21/2026 7:34:14 AM

Total Changes

798

Content

81 Replacements
53 Insertions
150 Deletions

Styling and Annotations

308 Styling
206 Annotations

Go to First Change (page 2)

Story Embeddings – Narrative-Focused Representations of Fictional Stories

Hans Ole Hatzel Universit"at Hamburg Language Technology Group [hans.ole.hatzel@uni-hamburg.
Chris Biemann Universit"at Hamburg Language Technology Group [chris.biemann@uni-hamburg.de]

Abstract

Story understanding requires capturing high-level narrative similarities that go beyond surface lexical overlap. We introduce StoryEmb, a set of embeddings fine-tuned specifically to capture narrative structure and thematic similarity in fictional stories. Our approach leverages weak supervision from naturally occurring retellings and adaptations of stories, encouraging representations that align narratively similar texts while remaining robust to differences in style, wording, and named entities. We evaluate StoryEmb on multiple narrative-focused retrieval and understanding tasks, including story and scene retrieval, movie remake matching, retelling identification, and the ROCStories cloze task. Across these benchmarks, StoryEmb consistently outperforms strong general-purpose embedding models, particularly in settings that require sensitivity to narrative structure rather than surface similarity. Qualitative and attribution analyses further suggest that StoryEmb places greater emphasis on plot-relevant content while de-emphasizing superficial cues. Our findings demonstrate that targeted fine-tuning with narrative supervision yields representations better suited for story-level semantic tasks.

1 Introduction

Narrative understanding is a field that has received much attention in the last few years. Various approaches have tested models either on narrative-based question answering tasks or performed intrinsic evaluations, such as narrative cloze evaluations, where models need to predict missing events in a sequence.

In this work, we seek to address the topic of story embeddings with a focus on narrative, meaning representations that prioritize the aspect of "what" is happening rather than the surface-level information of "how" it is being told. For example, a love story with a specific twist can be set in different settings (outer space or countryside), with a different cast (e.g., different names and some different traits for all characters), or in a shortened version, without fundamentally changing the narrative. After altering the story's final twist, the new narrative could still be considered similar without being identical.

Researchers in the ACL community have, in the context of fictional works, often used the terms narrative and story without a clear distinction (e.g., Chaturvedi et al., 2018; Chambers and Jurafsky, 2009). The field of narratology has a multitude of competing terms to offer, specifically to distinguish between the order of events as presented to the reader (commonly used terms are Syuzhet, Plot and Discours) and that of the actual happenings in the narrated world (commonly used terms are Fabula, Story and Histoire) (Kukkonen, 2019). In this work, we refer to the story as the entirety of the narration abstracted from the individual formulation, whereas we use narrative specifically to refer to the story's structure. Thus, a narrative could broadly be seen as the order and relationship of events in the story, but it does not include other information, such as the setting, tone, and style of the story.

This work presents a contrastive-learning-based approach for training story embeddings using a pre-existing dataset. We assume that any fictional text can be represented by its summary for our purposes of modeling the narrative. While there are various characteristics of a story that can not be gleaned from a summary, such as the style and mood of a text, the narrative is core to what is represented in a summary. Thus, summaries are the perfect testing ground for narrative embeddings, although an expansion to full texts in the future is desirable.

It has been observed that retellings of – specifically fairytales – have recently increasingly been published, with many retellings changing the setting to a modern-day one or introducing the representation of minorities (Goldman, 2023). As such, they represent a structurally similar story, with a new setting and limited alterations to the narrative. Other retellings, however, change the story significantly, sometimes merely retaining themes from the original work. On a limited scale, previous work has addressed the automatic identification of stories following the same plot (Glass, 2022). In this work, we consider this task as a possible application of story embeddings.

2 Related Work

A substantial line of work (e.g. Chambers and Jurafsky, 2008, 2009; Granroth-Wilding and Clark, 2016) has dealt with graph-based representations of narratives, specifically with predicting missing narrative triples and inferring schemas of commonly re-occurring narratives. Lee and Jung (2020) take what can be considered a hybrid approach, building explicit networks but using contextual vector representations rather than lexical items to represent triples. Similarly, using less contextual information, in prior work, we trained narrative triple embeddings based on narrative chains (Hatzel and Biemann, 2023). Following ever-increasing advancements in the field of language models and motivated by the information loss inherent to extracting narrative triples, this work seeks to apply a more distantly supervised approach to representing stories.

Our work builds on two previously released datasets (Hatzel and Biemann, 2024; Chaturvedi et al., 2018). Both datasets contain story summaries extracted from Wikipedia. Specifically, both seek to find different formulations of summaries for very similar stories. The movie remake dataset by Chaturvedi et al. (2018) contains a relatively small collection of summaries from multiple remakes of the same movie. In contrast, our previously released dataset, Tell-Me-Again (Hatzel and Biemann, 2024), collects summaries from multiple Wikipedia language versions of the same fictional work. The movie remake dataset only contains 266 summaries and is thus not suited for training, whereas Tell-Me-Again contains roughly 30,000 stories. Each story comes with up to five different summaries, originally extracted from multiple Wikipedia language versions and automatically translated into English. The dataset additionally comes with a pseudonymized variant, explicitly created for training models that do not focus on entity names. In this variant, entity names are replaced in each summary by alternatives in an internally consistent manner. These pseudonymized versions are created using rule-based replacement strategies on top of a model-based coreference resolution system.

ROCStories is a dataset for testing commonsense reasoning, first released in 2016 (Mostafazadeh et al., 2016) with the introduction of the Story Cloze Task. In the task, systems pick one of two sentences as the end of a five-sentence story. One choice is a logical conclusion to the story, but the other choice only matches in terms of vocabulary and is not a fitting conclusion to the story. As a result, humans can solve the Story Cloze Task perfectly, but at the time of publication, the best-performing system in an accompanying shared task reached only around 75

The creation of semantic sentence representations with large language models (LLMs) has recently gained much interest. While Wang et al. (2024) train embeddings from last-token hidden

states, it has been suggested that the causal attention mechanism in generative decoder-only models limits their effectiveness for embeddings (BehnamGhader et al., 2024). Alternatives have been proposed in the form of adding bidirectional attention back into existing models (BehnamGhader et al., 2024) or by duplicating the input sequence, thereby functionally allowing each token to attend to every other token (Springer et al., 2024). Ultimately, the new approaches were shown to be more training-sample-efficient but did not show real inference quality gains over the extensively finetuned E5 model by Wang et al. (2024).

Embedding approaches are typically focused on very short sequences of text, particularly individual sentences (Reimers and Gurevych, 2019; Ni et al., 2022). Doc2Vec (Le and Mikolov, 2014) is a static-embedding-based approach to document embeddings. While it was primarily evaluated on short segments, it does not have a limitation regarding the input size, a common constraint in transformer-based approaches.

The definition of what exactly constitutes narrative similarity has been addressed by Chen et al. (2022a) in their corresponding codebook (Chen et al., 2022b). In a pairwise similarity annotation task, they explicitly ask annotators to consider the narrative schemas and to ignore the specific names of entities, only considering their roles. They do not define an exact measure of how distances between schemas are determined, nor do they instruct annotators to write down explicit schemas. Despite these limitations, they achieve comparatively good inter-annotator agreement (0.69 Krippendorf's α) on narrative similarity of news articles.

3 Our Approach

Our model, called StoryEmb, is a causal language model whose last token representation is fine-tuned on similarity tasks using augmented data. Our model is trained to produce representations that are similar for multiple summaries of the same story. As a foundation model, we use Mistral-7B (Jiang et al., 2023a). Specifically, we use E5 (Wang et al., 2024), an adapter-finetuned variant, trained using synthetic data, for similarity modeling. As story similarity is a complex task, we assume that a more capable model would perform better; due to hardware constraints, we chose a 7B parameter model.

We train our model using Gradient Cache (Gao et al., 2021) to enable large batch sizes on limited hardware while reaching identical results to traditional similarity training. In training, we optimize for reducing the cosine similarity between pairs of summaries labeled as the same while maximizing the cosine similarity between those pairs that, by nature of belonging to different works, are implicitly labeled as different. Our approach follows Gao et al. (2021) in using contrastive MSE-loss for similarity training. We use a batch size of 1000 positive pairs and in-batch negatives. For the optimizer, we use Adam with a learning rate of 5×10^{-5} and perform early stopping on a subset of pseudonymized summaries from the Tell-Me-Again dataset. The training is limited to the adapter parameters and, as we are training based on their weights, we follow Wang et al. (2024) and use LoRA with rank $r = 16$ and $\alpha = 32$. While our training setup differs in various details (we use a different loss and do not employ hard negatives), the training can be considered a continued fine-tuning of E5 with a similar objective, just focusing on narrative similarity.

Our training data is sampled from the Tell-Me-Again dataset but limited to only summaries with a minimum of 10 and a maximum of 50 sentences in length. This is motivated by the desire to exclude (a) very short synopses and loglines on the low end and (b) documents that are too memory-demanding on the high end. The length limit could be subject to further experimentation in the future. We evaluate whether the data augmentation approach – replacing names with alternative ones in a consistent manner – proposed by Hatzel and Biemann (2024) can improve the performance

of a similarity model. To this end, we compare an augmented version of our model, trained on pseudonymized versions of the original summaries, and a non-augmented version, trained on the original summaries.

Following the E5 paper, we add a query prefix to each document. Through manual exploration on the development set, we selected the query, "Retrieve stories with a similar narrative to the given story: ". While many of the original applications of E5 follow an asymmetric setup where the query and the document are encoded using separate prompts, our prompt aligns well with one of their evaluation prompts: "Retrieve tweets that are semantically similar to the given tweet".

BehnamGhader et al. (2024) have recently introduced a more sample-efficient way, called LLM2Vec, to train LLMs for sentence representations. In preliminary experiments, we found, perhaps in part due to length limitations in training as a result of the full-attention setup, an LLM2Vec-based model to perform inferiorly to our model.

4 Experiments

After training, we perform several downstream task experiments to explore the capabilities and characteristics of our narrative embeddings. Three experiments test narrative retrieval capabilities (Section 4.1). We also perform an experiment focused on narrative understanding (Section 4.2). All our experiments in this paper are limited to English data. Recall that our training data consists of pairs of story summaries automatically translated from various languages to English.

4.1 Narrative Retrieval

Using four different tasks, we test if our embeddings can be used for retrieving narratively similar stories. All retrieval tasks are performed using embedding cosine distances.

For the initial three retrieval experiments, those with gold data available, we follow Chaturvedi et al. (2018) in using P@1 (precision at one), in other words accuracy for the most relevant result. Additionally, we introduce the P@N (precision at n) metric to allow for easy interpretation of the results. It measures the precision in the N-top results, where N is the number of gold items in the respective cluster. For reference, we also include the more established information retrieval metrics of MAP (mean average precision), NDCG (normalized discounted cumulative gain), and R-Precision (Manning et al., 2008).

4.1.1 In-Task Performance

In prior work (Hatzel and Biemann, 2024), we tested various existing models on pseudonymized and non-pseudonymized versions of the dataset, finding that all models, especially smaller ones, perform very poorly on the pseudonymized versions. In the existing publication, we attribute this to those models' reliance on entity names, showing that a bag-of-word system based only on entity mentions already performs well.

4.1.2 In-Domain Adaptation: Movie Remake Dataset

We expect retrieval performance on the movie remake task to be worse than on the Tell-Me-Again dataset, as one would expect summaries across remakes to show more variations than summaries sourced from various languages. This would align with our previous results (Hatzel and Biemann, 2024), where the best-performing model reached a P@1 of 64.4

4.1.3 Retellings

We collect a small set of summaries of works of fiction, each considered a retelling or a retelling's original. The collection methodology amounted to prompting ChatGPT for close retellings to limit the variations in the narrative.[1] The model was essentially used to suggest retelling relationships, and the list was subsequently checked for validity using manual web searches. While we considered other approaches, such as using existing lists of retellings, we decided, in part due to a lack of authoritative lists of this nature, to retrieve very commonly mentioned pairs using a language model instead. After discarding various suggestions that did not have English Wikipedia articles with plot summaries, we ended up with 13 clusters of retellings totaling 30 story summaries.

Retellings often change the story in major ways, more so than we would expect in a movie remake. We expect retellings to deviate more from each other than both multiple summaries of the same story and summaries of movie remakes. However, they may retain similar or identical character names, a characteristic that is not aligned with our pseudonymized training data. Given these characteristics, we initially anticipated that our model would find the retelling retrieval task more challenging than identifying movie remakes. We release the retelling dataset, including the full summaries, alongside our code, in a format matching that by Chaturvedi et al. (2018) for easy comparison.

4.1.4 Segment Retrieval

To generalize these findings to a broader story retrieval problem, we perform an annotation-based experiment, asking LLM judges and human annotators to rate the narrative similarity of text pairs. While a human-curated dataset of similar story pairs may also be desirable, we do not see a clear path to creating one. A human judgment of similarity relies on recalling a large set of stories, which is not generally achievable with annotators. So, our experiment instead relies on testing pairs of texts that the model considers to be very similar or dissimilar using human annotators. We follow Chen et al. (2022a) in broadly annotating for similarity in narrative schemas without making them explicit during annotation. A more precise definition of narrative similarity on the basis of schemas could be the subject of future work, but we do not consider it essential for this limited-scale experiment.

For this experiment, we select a moderately sized fiction dataset in which we expect to find frequent occurrences of similar scenes. We select a set of public-domain detective novels for this purpose.[2](https://www.gutenberg.org/ebooks/bookshelf/30) The novels are split into segments of no more than 2000 whitespace-separated tokens using a rule-based splitting solution.[3](https://github.com/umarbutler/semchunk) Said segments are subsequently summarized using LLaMA3's 70B[4](https://github.com/meta-llama/llama3) (at full 16-bit precision) variant with the prompt "Please summarize the following text in three sentences or less.". The resulting summaries are embedded using our StoryEmb model.

Initially, we remove all obviously similar pairs of summaries by discarding all pairs with a similarity higher than 0.3 according to MiniLM.[5](https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2) This ensures that duplicates that occur across documents in the dataset are not used as trivial examples of narrative similarity. We evaluate the similarity of the 50 most similar

[1] See Appendix B for the prompt and further details.
[2] [https://www.gutenberg.org/ebooks/bookshelf/30
[3] [https://github.com/umarbutler/semchunk
[4] [https://github.com/meta-llama/llama3
[5] [https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

segment pairs and 50 least-similar pairs in two setups: (a) first with an LLM judge and (b) with a human judge. For the latter, we sample just 10

4.2 Narrative Understanding: ROCStories

Finally, we perform an experiment aimed at validating the common-sense understanding of our model using the Story Cloze Task. In story cloze, given a common-sense story of four sentences, the system has to select the final fifth sentence of the story from two choices: an incoherent but surface-level-consistent ending and the correct and semantically coherent one. To test our embeddings, we take an unconventional approach to inference on this task, enabling evaluation without a classification head or similar techniques. We embed three components: the first four sentences of the story that we refer to as the anchor a and two variants of the entire five-sentence story with either the second or the first option: s_1 and s_2. Our system predicts the story that is closer to the anchor embedding. The intuition behind this is that a good story embedding already encodes expected outcomes, leading to a vastly different embedding for the incorrect, unfitting ending.

$$[\ \mathrm{m(a,\ s_1, s_2)} = \begin{cases} 1 & d(a, s_1) < d(a, s_2)\ 2 \\ d(a, s_1) \geq d(a, s_2) \end{cases}]$$

See Equation 1 for a more formal description, where d is an arbitrary distance measure, in our case cosine distance. For reference, we test not only our StoryEmb model but also other embedding models.

5 Results

As seen in Table 1, our StoryEmb model achieves state-of-the-art results on the Tell-Me-Again dataset, outperforming all other tested models in all but one metric. On the test set, our model, trained only using the augmented Wikipedia summaries, reaches a P@N of 65.89

We also test a pre-trained doc2vec model (Lau and Baldwin, 2016) as a more traditional baseline with no inherent length limitation. Outside of our own model's performance, it is interesting to see doc2vec outperform E5 by far on the pseudonymized version of the dataset; the static-embedding model exhibits no noticeable drop in performance from the standard to the pseudonymized setting (in fact, the results on the pseudonymized version are marginally better for all metrics). Upon consideration, this is not surprising as the static word embeddings in doc2vec may have a hard time with generic entity names, especially personal names.

While the performance increase on the pseudonymized texts is expected, it is surprising that, even for the non-pseudonymized texts, the model trained on augmented data performs better. As noted earlier, our model's training was stopped early based on the performance on the pseudonymized texts (for both model variants). The training finished after just three training steps (after seeing no improvements for two more steps), with each step taking roughly 1 h 40 min on two A100 GPUs. In fact, the unaugmented model continues to improve on the non-pseudonymized data afterward, presumably due to an ever-increasing focus on entity names as a shortcut to solving the task.

5.1 Movie Remakes

The results for the movie remake dataset listed in Table 2 are state-of-the-art for said dataset with a top P@1 score of 83.26

An interesting takeaway from the results on the movie remake dataset is a very pronounced drop in the performance of Sentence-T5 as compared to the Tell-Me-Again results. While the model

showed a P@N of 94.98 on the non-pseudonymized Tell-Me-Again data, its performance dropped by more than 17 points to 77.61

5.2 Retellings

The retrieval performance on the retelling dataset tests our model's capabilities in an alternative scenario with different requirements. On this dataset, Sentence-T5 outperforms our model by a considerable margin, reaching a P@1 of 70.0.

Interestingly, and despite the much smaller dataset size of only 30 rather than 266 summaries, our model's and the baselines' retrieval performance is much worse than on the movie remake dataset. This indicates that identifying retellings is a challenging task. At the same time, it is unclear if retellings are best identified using narrative features, given that they may only align with the story's themes. Our augmented models' underperformance may indicate that a name-focused approach is better suited to this task.

5.3 Scene Retrieval

Table 4 shows the narrative similarity ratings of our LLM judge and human annotator (the first author of this paper). The LLM judge favors the StoryEmb model when operating on the summaries of the retrieved segments, with the score increasing from 5.1 to 5.36 out of 10 when using our model instead of E5. This difference is much more pronounced when the LLM judge operates on the full segments instead. While our model is still rated at 5.36, the E5 model now only gets a score of 4.94. Our model also does better at retrieving dissimilar passages. We consider those passages retrieved by StoryEmb to have an average similarity of 3.6/10, in our annotations, whereas the segments retrieved by E5 score 4.6/10.

The passages retrieved by StoryEmb are significantly ($p < 0.05$) more narratively similar than those retrieved by E5. We use the Mann-Whitney U significance test (Mann and Whitney, 1947), as a normal distribution cannot be safely assumed. Remember that these scores are achieved on a set of segments prefiltered to remove obviously similar examples.

5.4 Story Cloze: ROCStories

Table 5 lists the results of our model on the ROCStories dev set. An accuracy of almost 90

The results show that an expected event in the story changes the embedding less than an unexpected one. Thus, this experiment indicates a high level of narrative understanding exhibited by our story embeddings.

6 Approximate Attribution

To inspect which aspects our StoryEmb model focuses on, we apply an attribution approach for sentence encoders (Moeller et al., 2024). The approach builds on the idea of integrated gradients (Sundararajan et al., 2017) and allows attributing similarity scores to individual input tokens.

We apply the method to pairs of summaries from the Tell-Me-Again dataset. Specifically, we analyze how much each token in a summary contributes to the cosine similarity between two summaries. We compare the attributions produced by our StoryEmb model to those of the base E5 model in order to identify differences in focus between the models.

Figure 1 shows attribution scores on individual tokens in the final layer of the StoryEmb model, displayed as a delta from the E5 model. Positive values indicate that StoryEmb assigns greater

importance to the token than E5, while negative values indicate a reduced importance compared to E5.

Our qualitative inspection suggests that StoryEmb assigns relatively higher importance to tokens that are central to the narrative structure, such as verbs describing key events and nouns referring to roles or actions relevant to the plot. In contrast, tokens corresponding to specific named entities, especially character names, often receive reduced attribution compared to E5.

To further quantify this observation, we analyze the average contribution to sentence similarity of selected named-entity and part-of-speech tags. Table 6 reports the average attribution scores grouped by tag for both E5 and StoryEmb. We observe that named entities receive lower relative importance in StoryEmb compared to E5, while certain verb and content-word categories show increased importance.

These findings align with our training objective and the use of pseudonymized data augmentation, which encourages the model to focus less on surface-level cues such as entity names and more on the structural and event-based aspects of narratives.

7 Qualitative Exploration

In addition to the quantitative experiments, we conduct a qualitative exploration of the similarities captured by our StoryEmb model. We manually inspect pairs of segment summaries that are considered highly similar by StoryEmb but not by the standard E5 model.

Figure 2 presents an example pair of segment summaries considered narratively similar by StoryEmb more so than the standard E5. In this example, the two segments describe structurally similar situations involving investigation and confrontation, despite differences in setting and specific details. While E5 assigns a comparatively lower similarity score to this pair, StoryEmb ranks them among the most similar segments.

Upon inspection, we find that StoryEmb appears to capture similarities in the sequence of events and the roles of characters rather than focusing primarily on overlapping vocabulary or shared named entities. The segments may differ substantially in surface realization, including names and specific contextual details, yet share a similar narrative progression.

At the same time, we also observe cases where StoryEmb retrieves segments that align more closely in terms of narrative structure than those retrieved by E5. These observations support our hypothesis that the model has learned to emphasize structural aspects of stories over superficial lexical cues.

However, qualitative analysis also reveals limitations. Some retrieved pairs share thematic elements rather than strict narrative structure, suggesting that the boundary between narrative similarity and thematic similarity remains fuzzy. Further work is needed to more precisely disentangle these notions and to refine the computational definition of narrative similarity.

8 Conclusion

In this work, we introduced StoryEmb, a model for creating narrative-focused embeddings of fictional stories. By leveraging contrastive learning on multiple summaries of the same story and employing pseudonymization as a data augmentation strategy, we trained embeddings that prioritize narrative structure over surface-level lexical cues.

Our experiments demonstrate strong performance across multiple retrieval tasks, including the Tell-Me-Again dataset and the movie remake dataset, where our model achieves state-of-the-art results. While performance on retellings remains more challenging, our findings suggest that

narrative-focused training leads to improved generalization across domains. Furthermore, our approach performs competitively on the ROCStories Story Cloze Task without task-specific training, indicating a degree of narrative understanding.

Attribution and qualitative analyses support the claim that StoryEmb places greater emphasis on plot-relevant tokens and reduces reliance on named entities. These findings indicate that targeted fine-tuning for narrative similarity can yield embeddings that better capture structural aspects of stories.

Overall, our results highlight the potential of narrative-focused embeddings for applications in story retrieval, recommendation systems, and computational literary analysis.

9 Future Work

Future work could explore extending our approach beyond summaries to full-length texts, enabling the modeling of richer narrative phenomena that are not fully captured in condensed plot descriptions. While summaries provide a practical and focused representation of narrative structure, incorporating full texts may allow for capturing additional subtleties in event progression and character development.

Another direction for future research lies in refining the notion of narrative similarity. A more formalized definition, potentially grounded in explicit narrative schemas, could improve both annotation consistency and model evaluation. Developing larger, human-annotated datasets specifically targeting narrative similarity would further facilitate progress in this area.

Additionally, investigating alternative training objectives, model architectures, or larger foundation models may yield further performance gains. Exploring multilingual extensions and cross-cultural narrative representations could also provide valuable insights into how narrative structures vary across languages and traditions.

Finally, potential applications such as story recommendation, adaptation analysis, and computational literary studies warrant deeper investigation, including user-centered evaluations and real-world deployment scenarios.

10 Limitations

Our work is subject to several limitations. First, our training and evaluation are restricted to English data. Although the Tell-Me-Again dataset is derived from multiple Wikipedia language versions, all summaries are automatically translated into English before training. As a result, our findings may not directly generalize to other languages without further experimentation.

Second, our model is trained and evaluated primarily on story summaries rather than full-length texts. While summaries capture core aspects of narrative structure, they omit stylistic, tonal, and other literary features that may be important for certain applications. Consequently, our embeddings may not fully represent all dimensions of a story.

Third, the retelling dataset used in our experiments is relatively small, containing only 30 summaries across 13 clusters. This limited size restricts the robustness of conclusions drawn from this evaluation and may not adequately reflect the diversity of retellings found in broader literary corpora.

Additionally, our qualitative and annotation-based evaluations are conducted on a limited scale. The human annotation in the scene retrieval experiment was performed by a single annotator, which may introduce bias and limits the generalizability of the findings. While we also employed an LLM judge, such evaluations are inherently dependent on the model's own biases and limitations.

Finally, our definition of narrative similarity remains somewhat informal and operationalized through contrastive learning on summary pairs. A more explicit and theoretically grounded formulation of narrative similarity could lead to clearer evaluation criteria and improved modeling approaches.

11 Ethical Considerations

Our work primarily uses publicly available data, specifically Wikipedia summaries and public-domain literary texts. We do not introduce new personal data, and our experiments are limited to fictional narratives. As such, we do not anticipate direct risks related to privacy or sensitive personal information.

However, the use of large language models for training and evaluation may inherit biases present in the underlying data and pretrained models. These biases could influence similarity judgments or retrieval behavior in ways that reflect existing societal biases. While our focus is on fictional stories, narrative content may still encode stereotypes or culturally specific assumptions.

The annotation-based evaluation involves human judgment of narrative similarity. Although limited in scope, such evaluations may reflect the annotator's subjective interpretation of similarity. Future work with a broader and more diverse group of annotators could help mitigate individual biases.

Finally, potential applications of narrative embeddings, such as recommender systems or automated literary analysis, should be deployed responsibly. Care should be taken to ensure transparency and to avoid reinforcing biased or narrow representations of narratives.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback and suggestions.

References

References

[1] BehnamGhader, Parinaz, Raghav Gupta, and Mohammad Taher Pilehvar. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. Preprint, arxiv:2404.05961.

[2] Chambers, Nathanael 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.

[3] Chambers, Nathanael and Dan Jurafsky. 2009. Unsupervised Learning of Narrative Schemas and Their Participants. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

[4] Chaturvedi, Snigdha, Dan Goldwasser, and Hal Daumé III. 2018. Story Comprehension for Predicting What Happens Next. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.

[5] Chen, Wenhu, Yu Su, and William Yang Wang. 2022a. Measuring Narrative Similarity for News Articles. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.

[6] Chen, Wenhu, Yu Su, and William Yang Wang. 2022b. Narrative Similarity Annotation Codebook. [ILLEGIBLE].

[7] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

[8] Glass, Michael. 2022. Detecting Plot Similarity in Fiction. [ILLEGIBLE].

[9] Goldman, Ari. 2023. The Rise of Modern Fairy Tale Retellings. Publishing Research Quarterly, 39(3):219–233.

[10] Granroth-Wilding, Mark and Stephen Clark. 2016. What Happens Next? Event Prediction Using a Compositional Neural Network Model. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, pages 2727–2733, Phoenix, Arizona, USA.

[11] Hatzel, Hans Ole and Chris Biemann. 2023. Narrative cloze as a training objective: Towards modeling stories using narrative chain embeddings. In Proceedings of the 5th Workshop on Narrative Understanding, pages 118–127, Toronto, Canada. Association for Computational Linguistics.

[12] Hatzel, Hans Ole 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, Turin, Italy. ELRA and ICCL.

[13] Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lelio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothee Lacroix, and William El Sayed. 2023a. Mistral 7B. Preprint, arxiv:2310.06825.

[14] Jiang, Yifan, Filip Ilievski, and Kaixin Ma. 2023b. Transferring Procedural Knowledge Across Commonsense Tasks. In 26th European Conference on Artificial Intelligence, September 30–October 4, 2023, Krakow, Poland, pages 1156–1163. IOS Press.

[15] Kukkonen, Karin. 2019. Plot. In The Living Handbook of Narratology. Hamburg: Hamburg University. http://www.lhn.uni-hamburg.de/article/plot](http://www.lhn.uni-hamburg.de/article/plot).

[16] Lau, Jey Han and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In Proceedings of the 1st Workshop on Representation Learning for NLP, pages 78–86, Berlin, Germany. Association for Computational Linguistics.

[17] Le, Quoc and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, pages 1188–1196, Beijing, China. PMLR.

[18] Lee, O-Joun and Jason J. Jung. 2020. Story embedding: Learning distributed representations of stories based on character networks. Artificial Intelligence, 281:103235.

[19] Mann, Henry B. and Donald R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, 18(1):50–60.

[20] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press.

[21] Moeller, Lucas, Dmitry Nikolaev, and Sebastian Padó. 2024. Approximate Attributions for Off-the-Shelf Siamese Transformers. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2059–2071, St. Julian's, Malta. Association for Computational Linguistics.

[22] Mostafazadeh, Nasrin, 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 the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 839–849, San Diego, California, USA. Association for Computational Linguistics.

[23] Ni, Jianmo, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5: Scalable Sentence Encoders from Pretrained Text-to-Text Models. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

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

[25] Springer, Jacob Mitchell, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition Improves Language Model Embeddings. Preprint, arxiv:2402.15449.

[26] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pages 3319–3328, Sydney, New South Wales, Australia. JMLR.org.

[27] Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. Preprint, arxiv:2401.00368.

[28] Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. In The Tenth International Conference on Learning Representations, Online. OpenReview.net.

A LLM Judge

We use GPT4-o, specifically gpt-4o-2024-05-13, in a multi-turn setup for similarity evaluation.

B Retelling Dataset

We collect a small set of summaries of works of fiction, each considered a retelling or a retelling's original. The collection methodology amounted to prompting ChatGPT for close retellings to limit the variations in the narrative. The model was essentially used to suggest retelling relationships, and the list was subsequently checked for validity using manual web searches. While we considered other approaches, such as using existing lists of retellings, we decided, in part due to a lack of authoritative lists of this nature, to retrieve very commonly mentioned pairs using a language model instead.

The prompt used to retrieve candidate retellings is as follows:

> [ILLEGIBLE]

After discarding various suggestions that did not have English Wikipedia articles with plot summaries, we ended up with 13 clusters of retellings totaling 30 story summaries. We release the retelling dataset, including the full summaries, alongside our code, in a format matching that by Chaturvedi et al. (2018) for easy comparison.

C Retelling Dataset Results

See Table 3 for the full table with all metrics on the retelling dataset.

D Data Code Availability

We release our code, model checkpoints, and the retelling dataset at [ILLEGIBLE].