Story Embeddings – Narrative-Focused Representations of Fictional Stories

Hans Ole Hatzel
Universität Hamburg
Language Technology Group
hans.ole.hatzel@uni-hamburg.de

Chris Biemann
Universität Hamburg
Language Technology Group
chris.biemann@uni-hamburg.de

Abstract

We present a novel approach to modeling fictional narratives. The proposed model creates embeddings that represent a story such that similar narratives, that is, reformulations of the same story, will result in similar embeddings. We showcase the prowess of our narrative-focused embeddings on various datasets, exhibiting state-of-the-art performance on multiple retrieval tasks. The embeddings also show promising results on a narrative understanding task. Additionally, we perform an annotation-based evaluation to validate that our introduced computational notion of narrative similarity aligns with human perception. The approach can help to explore vast datasets of stories, with potential applications in recommender systems and in the computational analysis of literature.

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.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5931–5943
November 12-16, 2024

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% accuracy. The original task formulation did not allow for supervised learning, providing only complete five-sentence stories without two choices as training data.

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 reprentations. 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% on the remake dataset but reached 90.5% on the Tell-Me-Again dataset (in a setup where the Tell-Me-Again dataset was subsampled to replicate the movie remake dataset's distribution). As for the Tell-Me-Again dataset, we test on both the original summary and the pseudonymized version, with two model variants trained on either variant of the dataset.

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 retelling the 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] The novels are split into segments of no more than 2000 whitespace-separated tokens using a rule-based splitting solution.[3] Said segments are subsequently summarized using LLaMA3's 70B[4] (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] This ensures that duplicates that occur across documents

[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

in the dataset are not used as trivial examples of narrative similarity. We evaluate the similarity of the 50 most similar 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% of segments, using the same similarity ranks for both models (sampling from the same ranks in terms of similarity in the pseudonymized and the standard E5 model with a task prefix). Judges are asked to rate the similarity of segments on a scale of 1-10. The LLM judge evaluates our embeddings in two scenarios based on the segment's original full text and its automatically generated summary. For time reasons, the human judge only operates on automatically generated summaries. For the judge model, we use GPT4-o in a two-turn setup; for further details on the LLM judge setup, see Appendix A.

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.

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

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% in the pseudonymized setup on a dataset of almost 10,000 story summaries. That is to say, over all retrieved summaries, which correspond to the number of gold reformulations for each summary, 65.89% of them are correct. This is a pronounced drop compared to the 85.90% on the original texts, but compared to other models, the drop is much smaller. This is also true for the most competitive model, Sentence-T5 in its XXL variant (Ni et al., 2022), which marginally outperforms StoryEmb in P@1 on the non-pseudonymized dataset. Sentence-T5 reaches a P@1 of 94.98%, while StoryEmb only reaches 94.64% on the original data. On the pseudonymized data, however, Sentence-T5 only reaches a P@1 of 67.28%, while our approach yields 82.6%.

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%, improving by more than 20 points over the original story-kernel approach by Chaturvedi et al. (2018). On this dataset, we also outperform Sentence-T5 by a considerable margin, reaching 80.28%, an almost 7-point improvement over their result of 73.35% in terms of P@N. Again, we can clearly observe the positive effects of the pseudonymization data augmentation. We also provide results for the unaugmented StoryEmb model trained for two more steps, thereby almost doubling the fine-tuning data. Yet, despite the additional training data, our non-augmented model on the non-pseudonymized dataset does not meaningfully improve. Additional training only improves the P@1 score of 63.09% by just over .2 points to 63.30%, clearly demonstrating the effectiveness of the data augmentation approach. Both versions of our model substantially outperform the base E5 model, an effect that we attribute to domain adaptation, including an adaptation to longer documents.

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% on the movie remake dataset, while our augmented StoryEmb model lost less than 6 points on the same metric (85.9% to 80.28%) across the datasets. Initially, we suspected that this may be caused by a case of training data leakage, with T5 having incorporated Wikipedia cross-language version training. This could not be confirmed as, when limiting evaluation to works from 2022 or 2023, after the release of Sentence-T5, it actually performed comparatively even better, reaching a P@1 of 96%, where our augmented model only reached 83%. This points to a semantic difference in remakes compared to language versions that StoryEmb captures better as the explanation, demonstrating superior generalization capabilities for narrative retrieval in StoryEmb.

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%, while our best-performing model-variant, the unaugmented model, reaches 60.0%. Our model still handily outperforms vanilla E5 at a P@1 of 16.67%. Additionally, despite its great performance on previous tasks, the model trained on augmented data now underperforms as compared to the unaugmented version. Given the dataset's limited size, we expect that an entity-focused approach works better. We expect that names serve as easy disambiguators in a smaller dataset but lose discriminative performance as the number of samples grows. To test this hypothesis, we add the summaries from the movie remake dataset as distractors to the task. In this setup, we see the margin by which T5 overperforms shrink considerably, especially for the P@N metric, where the best Sentence-T5 model now only outperforms our unaugmented model by roughly 2 points, with a score of 57.69% as compared to 59.62%. See Appendix C for the full table with all metrics.

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 LLM judgments on the full-text segments

Name	Pseudonymized					Non-Pseudonymized				
	MAP	NDCG	R-precision	P@1	P@N	MAP	NDCG	R-precision	P@1	P@N
Doc2Vec	38.02	53.83	35.19	51.30	34.96	37.99	53.81	35.13	51.27	34.91
E5	22.07	38.57	20.54	33.20	20.87	54.67	68.38	51.51	73.65	51.87
StoryEmb + aug	**72.83**	**82.47**	**67.32**	**82.60**	**65.89**	**90.05**	**94.10**	**86.54**	94.64	**85.90**
StoryEmb	36.06	52.71	32.38	47.63	31.25	73.43	82.80	68.36	83.78	68.49
Sentence-$T5_{XXL}$	54.67	68.48	50.09	67.28	48.16	89.21	93.48	86.15	**94.98**	85.61

Table 1: Retrieval performance on the Tell-Me-Again test set by Hatzel and Biemann (2024), with and without their anonymization strategy.

Name	Pseudonymized					Non-Pseudonymized				
	MAP	NDCG	R-precision	P@1	P@N	MAP	NDCG	R-precision	P@1	P@N
Doc2Vec	44.89	56.48	35.59	38.41	35.51	52.36	62.65	43.67	46.78	43.65
E5	26.00	39.62	20.03	22.75	20.15	51.65	61.19	45.78	49.36	45.39
StoryEmb + aug	**78.18**	**83.30**	**72.39**	**75.11**	**72.44**	**84.67**	**88.45**	**80.51**	**83.26**	**80.28**
StoryEmb	47.15	58.55	37.77	40.56	37.94	66.18	73.97	58.62	63.09	58.17
StoryEmb + 2 steps	49.87	60.96	40.84	44.42	40.81	66.37	74.04	59.73	63.30	59.26
Sentence-$T5_{XXL}$	56.89	66.55	48.50	51.93	48.58	79.21	84.24	73.75	76.61	73.35
Chaturvedi	-	-	-	-	-	-	-	-	63.7	-

Table 2: Test set retrieval performance on the dataset by Chaturvedi et al. (2018), with and without the anonymization strategy by Hatzel and Biemann (2024) applied to the dataset. "+2 steps" denotes two additional steps of training.

Model Name	R-precision	P@1	P@N
doc2vec	18.33	16.67	17.31
E5	35.00	36.67	36.54
StoryEmb + aug	50.56	56.67	51.92
StoryEmb	58.33	60.00	59.62
Sentence-$T5_{XXL}$	**62.78**	**70.00**	**67.31**
+ Movie Remakes as Distractors			
StoryEmb + aug	48.33	50.00	51.92
StoryEmb	53.33	50.00	57.69
Sentence-$T5_{XXL}$	**55.00**	**60.00**	**59.62**

Table 3: Retrieval performance on retelling dataset introduced in Section 4.1.3, optionally with the movie remakes added as distractors.

	LLM Judge				Annotator	
	Summary		Full-Text		Summary	
Subset	E5	Ours	E5	Ours	E5	Ours
Top	5.1	5.36	4.94	5.36	5.8	6.6
Bottom	4.06	3.84	4.22	3.72	4.6	3.6

Table 4: Mean narrative similarity score on a scale of 1-10 in top vs. bottom ranked scenes in terms of similarity as judged by an LLM judge or an annotator, after removing obvious duplicates. The first author performed the annotations.

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% shows that our embedding approach to ROCStories performs well with the StoryEmb model despite a lack of task-specific training. While our approach does not achieve overall state-of-the-art performance, it outperforms a few-shot prompting approach on GPT-3 with only 7B parameters. At the same time, Sentence-T5, an otherwise strong model, does not exhibit great performance. Note that our results are obtained on the development set. Recall that we do not perform specific training for this task; We neither train on next-sentence prediction nor use training data similar to the story cloze dataset.

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 (Sun-

System	Setup	Accuracy
Random	-	50.0
E5$^{\Delta}$	zero-shot	78.5
T5-XXL$^{\Delta}$	zero-shot	85.8
StoryEmb + aug$^{\Delta}$	zero-shot	89.2
StoryEmb$^{\Delta}$	zero-shot	89.4
GPT-3	zero-shot	83.2
GPT-3	few-shot	87.7
FLAN 137B	zero-shot	93.3
FLAN 137B	few-shot	94.7
RoBERTa	supervised	97.9

Table 5: We list the accuracy at picking the correct story ending from two options on the ROCStories dataset. The superscript$^{\Delta}$ denotes that the embedding distance approach outlined in Section 4.2 is used and evaluated on the development set. The GPT-3 and FLAN results are taken from Wei et al. (2022), and the supervised RoBERTa result is taken from Jiang et al. (2023b).

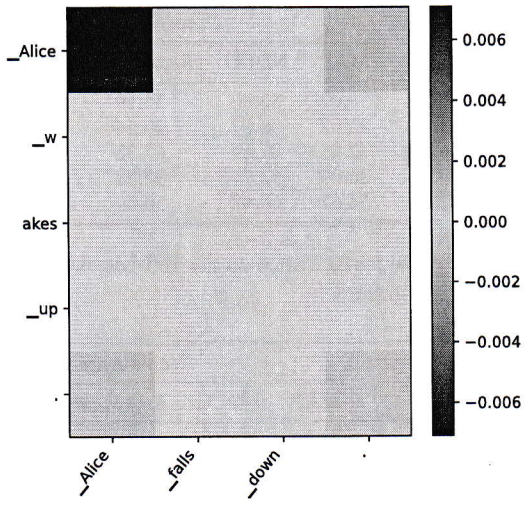

Figure 1: Attribution scores on individual tokens in the final layer of our StoryEmb model are shown as a delta from the E5 model. Negative scores indicate less contribution to the similarity in the StoryEmb model. In the example, it seems clear that less emphasis is placed on named entities.

dararajan et al., 2017), extending the concept to Siamese networks, specifically sentence encoders. In essence, the approach samples gradients along an interpolation from a semantically neutral sequence to the analyzed sequence, identifying features that the output is sensitive to. The result is a token-token matrix across two input sequences, signifying the contribution of any two terms to the overall similarity of the two sequences.

Moeller et al. (2024) operate only on models that use average pooling across tokens. We adapt their implementation to work with decoder-only models and use 50 interpolation steps. E5 employs last-token pooling, using the last token's hidden state as a sentence representation. As a result, the attribution scores are muddled: information needs to flow to the last token, leading to the majority of the similarity being explained by changes in the last token's representation. In an effort to get interpretable attribution scores despite the pooling approach, we display the delta in attribution scores from the E5 model to our StoryEmb model instead.

Figure 1 illustrates that our augmented-data approach does, in fact, place less emphasis on named entities than the vanilla E5 model. We compare similarity across the two sentences "Alice wakes up." and "Alice falls down.". Generally, negative values in an attribution indicate that two specific tokens interact to reduce the similarity of the two sentences. The attribution scores Figure 1 are computed by deducting the E5 attribution values from the StoryEmb attribution values. Thus, a negative value means our StoryEmb model places comparatively less emphasis on said value. In this case, our fine-tuned model places much less importance on the name "Alice" than the original E5 model does.

To generalize from the single example, we collect attribution scores for 50 random sentences from the STS benchmark dataset's test set (Cer et al., 2017). We collect average attribution scores for part-of-speech and named-entity tags, showing the results in Table 6. One can see the expected effect from our training on the part of speech tags; the model places much less emphasis on proper nouns (PROPN), while verbs (VERB) contribute slightly more to similarity scores. This effect is also visible for the named entity tags, with persons (PERS) and organizations (ORGS) being considered less relevant. Using the named-entity tags, however, we can also observe an unwanted side effect of the data augmentation; as dates are not removed by the data augmentation, they can still serve as a shortcut for solving the task, and our StoryEmb model prioritizes them.

We have restricted the analysis to single sentences as it is computationally expensive and could not feasibly be performed on entire stories.

7 Qualitative Exploration

In general, the manual evaluation of similarity has the issue that many datasets contain some form of duplicates, leading to the issue that trivially similar narratives are retrieved first. We explored similar

Tag-Kind	Tag	Similarity Contribution		
		E5	StoryEmb	Delta
PoS	PROPN	.0005	.0003	-.0002
	CCONJ	.0001	0	-.0001
	NOUN	.0002	.0002	0
	VERB	.0001	.0002	.0001
	PART	.0001	.0002	.0001
Entity	ORG	.0016	.0005	-.0011
	PERSON	.0004	.0001	-.0003
	GPE	.0005	.0004	-.0001
	DATE	.0028	.0043	.0015

Table 6: The average contribution to sentence similarity of selected named-entity and parts-of-speech tags was analyzed on layer 31 of the E5 and StoryEmb models. The statistics exclude our task prefix.

segments, using the approach from Section 4.1.4, excluding trivially similar texts. We found many instances of meaningfully narratively similar texts being retrieved by StoryEmb. In Figure 2, we show a pair of segment summaries from the detective novel dataset, with the segment from "The Triumphs of Eugene Valmont" being considered the 17th closest, by StoryEmb, to the "In the Fog" segment. E5, meanwhile, only considers it to be the 221st closest segment. The two segments share a very similar narrative of a necklace theft with a subsequent arrest but lack any shared named entities.

> The narrator, a Queen's Messenger, was traveling from Paris to Marseilles with a valuable diamond necklace belonging to the Queen of England. During the journey, he was distracted by a charming woman who stole the necklace from his bag while pretending to be friendly and conversational. When the narrator discovered the theft, he rushed to the police station and used his credentials to demand assistance in capturing the thief, explaining that the successful recovery of the necklace would bring great rewards and gratitude from three powerful nations.
>
> *Summarized segment from "In the Fog" by Richard Harding Davis*

> A private detective from London explains how he became involved in a case involving a stolen necklace and a man who jumped overboard from a yacht. The detective followed the thief and ended up on the yacht, where he claims the man may have escaped with the jewels. Despite his story, the Englishman is arrested and held for three weeks until a letter arrives from New York, revealing that the missing Frenchman, Martin Dubois, is alive and blaming the police for any harm that may have come to him.
>
> *Summarized segment from "The Triumphs of Eugene Valmont" by Robert Barr*

Figure 2: Our model considers these two texts much more similar than the standard E5.

8 Conclusion

In this work, we presented an approach to creating embeddings that represent stories, specifically their narratives. Our StoryEmb model, trained on the Tell-Me-Again dataset, shows state-of-the-art performance on both the corresponding test split and on the movie remake dataset (Chaturvedi et al., 2018), far outperforming both recent LLM-based models and the static-embedding baseline. We demonstrate the model's retrieval capability in practice on summaries of passages of literary texts. Further, we explore the retrieval of retellings on a novel small-scale dataset, opening up another potential application avenue.

On the movie remake dataset, we clearly demonstrate the effectiveness of the Tell-Me-Again data augmentation approach, producing better results on both pseudonymized and non-pseudonymized versions of the texts. Similarity score attribution indicates that the data augmentation techniques employed have the desired effect: making the model place less emphasis on names.

Our StoryEmb model also performs strongly on the StoryCloze task, indicating some level of narrative understanding.

9 Future Work

In the future, we expect to explore application fields for story embeddings further. Outside of employing larger foundation models, the contrastive learning approach can potentially be improved. We assume that one may also, at the cost of compute resources, be able to create a better model by not using a pretrained similarity model but instead starting from a plain language model.

Similarly, there is potential to generate better training data by (a) improving the pseudonymized versions or (b) creating new summaries, potentially by combining multiple existing ones using LLM prompting.

10 Limitations

Regarding training data leakage, we can assume that the foundation model has seen all Wikipedia

summaries. We do not expect this to be a substantial issue as we expect the different language versions to individually be trained on, as evidenced by relatively poor performance without further training. However, it is possible that the data augmentation strategy has limited success with entity names being inferred from the unredacted text seen in training.

It is possible that, given further hyperparameter tuning, the results could be noticeably improved. Contrastive learning, especially in the image space, has seen many optimizations. Due to resource constraints, this work was out of scope for this study. In initial experiments, we did not succeed with batch-sampling techniques, but it can be assumed that further exploration could yield improvements.

The representations we present lack interpretability compared to schema-based approaches to narrative modeling. At present, we cannot confidently identify which aspects of a story and its narrative are captured in our embeddings, and more work is required to understand exactly which information is captured by StoryEmb.

Our ROCStories results were obtained on the development set as the test set is privately held. We contacted the original authors and researchers who recently reported results on the dataset but were unable to get our predictions for the private test set scored. We have no reason to believe our performance on the test set would be worse.

11 Ethical Considerations

We do not see any major ethical problems. The data augmentation strategy in the original Tell-Me-Again dataset picks names based on US census statistics, potentially contributing to a system that may be regionally and culturally biased. This limitation is inherent to many NLP systems and should be addressed before approaches like this are used productively.

Acknowledgements

The work was supported by the German Research Foundation (DFG) under grant BI 1544/11-2 as part of the project "Unitizing Plot to Advance Analysis of Narrative Structure (PLANS)".

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. *Preprint*, arxiv:2404.05961.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: Human Language Technologies*, pages 789–797, Columbus, Ohio, USA. Association for Computational Linguistics.

Nathanael Chambers 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 of the AFNLP: Volume 2 - ACL-IJCNLP '09*, volume 2, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

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, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022a. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, Washington, USA. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Q. Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw A. Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022b. SemEval-2022 Task 8: Multilingual news article similarity. `https://doi.org/10.5281/zenodo.6507872`.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

Grant Glass. 2022. An Adaptive Methodology: Machine Learning and Literary Adaptation. In *Digital Humanities 2022 Combined Abstracts*, Tokyo, Japan. DH2022 Local Organizing Committee.

Melanie Goldman. 2023. The Rise of Fairytale Retellings in Publishing. *Publishing Research Quarterly*, 39(3):219–233.

Mark Granroth-Wilding 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.

Hans Ole Hatzel 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.

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, Turin, Italy. ELRA and ICCL.

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

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

Karin Kukkonen. 2019. Plot. In *The Living Handbook of Narratology*. Hamburg: Hamburg University. `http://www.lhn.uni-hamburg.de/article/plot`.

Jey Han Lau 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.

Quoc Le 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.

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

Henry B. Mann 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.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Evaluation in information retrieval. In *Introduction to Information Retrieval*, pages 151–175. Cambridge University Press.

Lucas Moeller, 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.

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 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.

Jianmo Ni, 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.

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

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

Mukund Sundararajan, 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.

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

Jason Wei, 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. The model is first asked to describe similarities and differences in the narratives of both segments, and only in a second step is it asked to submit a rating. Our conversation template looks as follows:

System: "You are a helpful assistant specializing in fictional texts and their narrative."

User: "In which respects, particularly focused on the narrative, are the following two stories similar and dissimilar? Focus on the structure of the story. Do not focus on specific names or places.

Assistant: ...

User: "Now based on your assessment, rate the similarity of the two stories on a scale of 1-10 where 10 is very similar. Please use the format X/10."

Assistant: ...

The sequence "..." denotes that we let the LLM generate the response. We use a temperature of 0 for both generation steps.

B Retelling Dataset

We prompt ChatGPT with the following prompt: ""Little Fuzzy" is a modern retelling of "Fuzzy Nation" that is relatively close in terms of story it can thus be considered a close retelling. What are some other pairs of close retellings?"

Note that we did not see a large variation of retellings produced on variations of the prompt, with many pairs frequently reoccurring even when explicitly asking for distant or far remakes. For this reason, we decided not to build two splits of the dataset with far and close remakes.

C Retelling Dataset Results

See Table 7 for our detailed results on the retelling dataset.

D Data & Code Availability

Our code and models are available online: `https://github.com/uhh-lt/story-emb`.

Model Name	MAP	NDCG	R-precision	P@1	P@N
doc2vec	29.99	47.05	18.33	16.67	17.31
E5	42.72	56.97	35.00	36.67	36.54
StoryEmb + aug	61.85	72.33	50.56	56.67	51.92
StoryEmb	63.55	73.27	58.33	60.00	59.62
Sentence-$T5_{XXL}$	**69.96**	**78.70**	**62.78**	**70.00**	**67.31**
+ Movie Remakes as Distractors					
StoryEmb + aug	56.43	66.53	48.33	50.00	51.92
StoryEmb	55.72	65.08	53.33	50.00	57.69
Sentence-$T5_{XXL}$	59.15	68.28	55.00	60.00	59.62

Table 7: Retrieval performance on retelling dataset introduced in Section 4.1.3