Narrative Similarity Detection Using Character Roles

Alexandra Barancová[1,*], Giovanni Sileno[1]

[1]*Informatics Institute, Faculty of Science, University of Amsterdam, the Netherlands*

Abstract

A growing domain of scholarship foregrounds the relevance of narrative information for text analysis. Current approaches to text similarity detection lean primarily on document-level embeddings. However, these methods are unable to capture structured narrative information in order to allow document comparison that explicitly takes into account narrative features. In this study, we examine how structured narrative elements, extracted using LLMs, can complement document-level embeddings for narrative similarity detection. To test the proposed pipeline, we assess whether character role information improves performance on similar story retrieval for medium- and long-form narratives. Focusing on a dataset of movie summaries, our results show a slight increase in performance when integrating structured character role features to measure story similarity. The framework can be easily extended to integrate further narrative components.

Keywords

computational narrative understanding, narrative similarity detection, character role extraction, film summaries

1. Introduction

The importance of stories for human sense-making cannot be overstated, and, for this reason, comparing stories is a relevant task for a range of applications. The interest in handling narratives by computational means is a longstanding topic in AI, starting from symbolic traditions [1], which has recently found its way into quantitative, data-driven methods, like clustering [2] and recommender systems [3], to name a few. However, the inherent multidimensionality of narratives makes this a complex task. Not only is it challenging to determine which features are important to consider for a given story in a particular context, but it is also often difficult to find ways to meaningfully combine the various dimensions.

This study elaborates on some of these challenges by focusing on two interconnected questions. (1) We examine how structured narrative elements, extracted using large language models (LLMs), can complement document-level embeddings for narrative similarity detection. Based on current practices observed in the literature, we settle on a simple general pipeline for experimentation. (2) To demonstrate its potential value, we focus specifically on *character role* information and assess whether it improves performance on similar story retrieval for medium- and long-form narratives.

Previous computational approaches demonstrate the efficacy of *entity data* for story comparison [4, 5]. However, to enable comparisons that reflect deeper narrative features rather than surface-level lexical cues, we propose to abstract concrete named entities into character roles that capture aspects of their narrative function. We hypothesize that this abstraction may offer a more generalisable basis for comparison, for example, between stories whose heroes do not share the same names or inhabit the same (fictional) universe. To test this hypothesis, we conduct our experiments on the *Tell me Again!* dataset, which collects several descriptions of the same film, each sourced in different languages on Wikipedia and translated into English [5].

Our method involves combining structured narrative elements with document embeddings to assess the similarity of two narrative texts. More concretely, our contribution extends Hatzel and Biemann's [5] baseline approach for narrative similarity detection using the character role schema proposed by Hobson et al. [6]. It is worth emphasising that the proposed pipeline can easily be adapted to employ alternative or additional schemas; this paper primarily serves as a proof-of-concept application. Unlike contemporary

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'26 Workshop, Delft (The Netherlands), 29-March-2026

*Corresponding author.

a.barancova@uva.nl (A. Barancová); g.sileno@uva.nl (G. Sileno)

methods that largely depend on document embeddings, this approach opens a space of encounter between structure-oriented research (in narratology and computational ontologies) and the benefits of LLM-based tools.

2. Related work

Here we outline previous work that informs our study: firstly, background to the task of narrative similarity detection; and secondly, ways in which characters have been theorised, modelled and studied, including in recent LLM-driven approaches.

2.1. Modelling narrative similarity

Our research is positioned within the growing domain of computational narrative understanding — for an overview, see for example [7]. Much of the work in computational narrative understanding seeks to explore and characterise various features of narratives; one application is that these can serve as grounds for comparison. In developing models, formal approaches often build on literary theory and narratology, e.g. [8, 9, 10], while cognitivist approaches tend to employ empirical and participatory methods to uncover dimensions relevant to human assessment of narratives, e.g. [11, 12, 13, 14]. The latter point to the discrepancies between the two, showing, for example, that non-experts tend to consider non-structural data [11], multiple different levels of story representation at once [12], or simply dimensions that are often omitted in expert accounts, like genre and style [13].

Although the contours for story comparison have been widely theorised, tackling this task computationally has been picked up relatively recently. Framed as an *information retrieval* problem, narrative similarity detection becomes functional to identify instances of similar narratives from a collection of narrative texts [4]. Given the lack of data with ground truth labels, previous work has proposed to make use of proxies, assuming, for instance, that summaries of movie remakes [4] and different summaries of the same film [5] can be considered instances of the same story.

2.2. Character roles in narrative research

Characters play an important role for narrative understanding in both formal and cognitivist approaches. Various frameworks have been developed and adapted for (computational) story analysis. For generalisable approaches to narrative similarity detection, such character models are crucial; although prior recent work demonstrates the efficacy of inference based on low level lexical features like named entities [4, 5], these examples also illustrate the need for character models that capture deeper, narrative properties of personas, in order to adequately address narrative semantics.

Vladimir Propp's structuralist analysis of Russian folk tales [8] offers one popular taxonomy of character roles (*dramatis personae*): hero, villain, donor, helper, princess/prize, false hero, dispatcher. Recent work has explored unsupervised methods for labelling these roles in narratives [15] and experimented with extending this to automatically identifying Propp's character functions [16]. The Hero-Villain-Victim (HVV) framework is an abstracted narrative role taxonomy that has been more widely explored and adapted across other domains including news [17, 6], memes [18], political speeches, and films [19]. Other works have also made use of altogether different narrative frameworks — e.g. adapting Greimas' actant theory [9] to compare roles in news narratives [20] — or more fine-grained, domain-specific taxonomies. Bamman et al. [21], for example, developed an approach to learn personas from a latent space of film summaries, using which they identified 72 character trope classes that others, like Chu et al. [22], have since used for character labelling.

At a technical level, the majority of recent examples listed here have been experimenting with using various LLMs for feature extraction or classification tasks based on preset narrative schemas [23, 16, 19, 20]. In addition to their classification capabilities, generative language models allow for more open labelling paradigms. Hobson et al. [6], for example, propose a taxonomy-free approach to character role labelling that we build on in this study, which queries an LLM for a modifier and primary label (e.g.

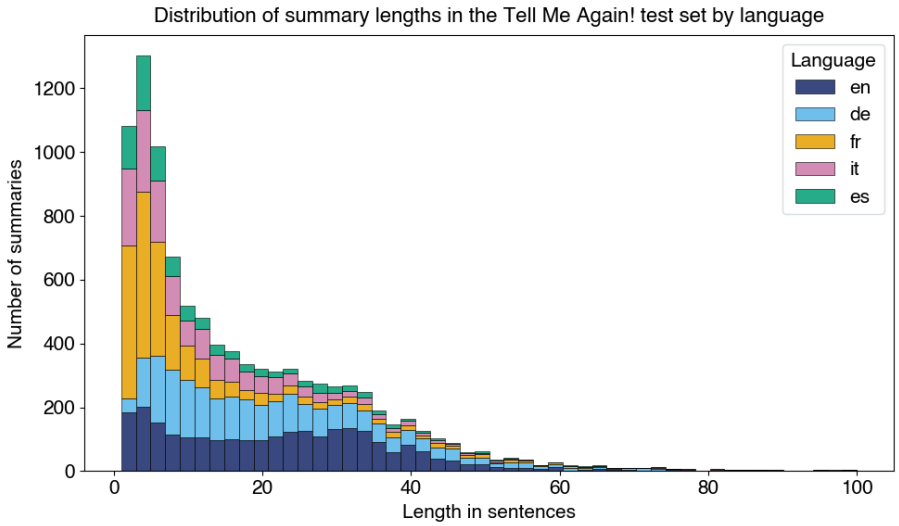

Figure 1: Distribution of the summary lengths in the *Tell Me Again!* test set (n=9778), given by the number of sentences in each text, grouped by the source language. This overview zooms in on summaries with ≤100 sentences — 56 longer summaries are left out for better legibility.

'Innocent Protestor'). This promises more flexible, rich, auditable, and contextually-tailored methods, while opening up new challenges, such as, how to narrow the space of possibility in order to make meaningful comparisons, or how to account for and understand the biases that such models introduce.

3. Methodology

The literature presented in Section 2.1 highlights that it is unclear which structural dimensions contribute to judgments of similarity. Yet, LLMs themselves, by means of their embedded space, may provide a geometric space in which to measure distance, even between compressed versions of the input. Our pipeline aims to investigate this intuition. Below we outline our data, experimental setup and evaluation approach.[1]

3.1. Data

We use the *Tell me again!* dataset [5], which consists of a total of 96,831 individual summaries across 29,505 stories. The dataset was assembled specifically for the task narrative similarity detection, offering a proxy for ground truth labels; assuming that different summaries of the same film recount the same story. The dataset combines summaries from Wikipedia pages in different languages for the 29,505 films, namely, the English, German, Italian, French, and Spanish versions. Non-English summaries are translated into English. The dataset also offers an anonymised version of each summary, where named entities are replaced with strings of the form 'Entity A', 'Location B', or 'Organization C' according to their respective entity tags. Names of people are replaced with randomly sampled names instead; these are associated with the same gender and follow the name distribution in US-census data.[2]

In this study, we focus on a test set consisting of 2,951 unique stories and 9,778 summaries, with a mean of 3.31 summaries per story. This set is filtered to exclude (near-)duplicates, resulting, for example, from direct translation between the different Wikipedia language versions or a shared external source.[3] Figure 1 presents the distribution of the number of sentences in the test set, with mean=17.4 and median=12.

[1]Our code is available at https://github.com/abaran3/character-roles

[2]All named entities were detected using Flair and a coreference resolution model.

[3]Filtering is done based on BERTScore similarity measures, which are provided for each story cluster as part of the dataset. We drop summaries above a similarity score of 0.6, the same threshold as Hatzel and Biemann [5]. De-duplication takes into

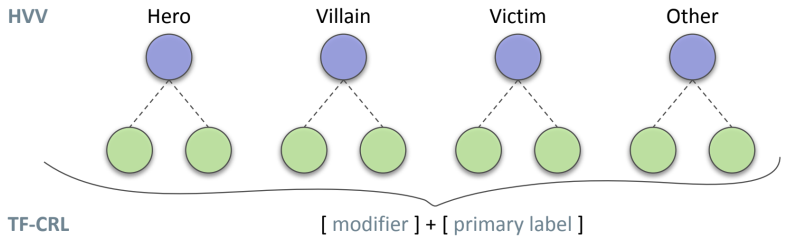

Figure 2: Illustration of the two-level character role model used to represent each film summary. It combines the Hero-Villain-Victim (HVV) framework and the taxonomy-free character role (TF-CRL) schema introduced by Hobson et al. [6]. The number of character nodes (green) is not fixed.

3.2. Experimental setup

In our experiments, we compare the performance of four sentence encoder models using several different representations of the summaries for the narrative similarity task. An initial exploration looks at the full test set of English language summaries as well as their anonymised versions, across different summary lengths. We also extract character roles for each summary in two steps, narrowing our focus to medium-long summaries with length $5 \leq n < 60$ sentences.[4] Below we outline these steps in detail, focusing first on the character role extraction process, and then the embedding generation, which was conducted for summaries as well as the character role information for evaluation.

3.2.1. Entity detection

The first step in our character role extraction process entails identifying relevant entities for each summary in the test set using named entity recognition and coreference resolution with spaCy[5] and Coreferee[6]. Like Hobson et al. [6], we define a character as any 'animate object that is important to the plot' [24] and include individual and collective entities.[7] Using this entity detection step allows us to isolate the character role labelling task (instead of e.g. using an LLM to identify *and* label characters); we opted for this approach for better traceability and repeatability.

3.2.2. Character role labelling

The above preprocessing step produces a character list for each summary, which we then use to query an LLM (gemma3:4b) for labels following the integrated HVV and taxonomy-free character role model illustrated in Figure 2. Gemma3:4b is a lightweight multimodal model with open weights [25] that worked efficiently in our preliminary tests. We use parameters *temperature*=0.05, *top_p*=0.95, *repeat_penalty*=1.1 and query the LLM using Ollama with a Pydantic model to ensure that the outputs conform to our character role schema. This allows us to extract a complete set of character role labels in a single zero-shot pass for each summary.[8]

As illustrated in Figure 2, the character-focused narrative model consists of two levels of granularity: a closed label based on the Hero-Villain-Victim framework and an open, taxonomy-free label that describes each character's role in a short noun phrase. Although we considered a more fine-grained closed taxonomy, we adopted HVV in the current setup. Previous work has shown that infrequent and highly distinct Proppian roles (e.g. dispatcher or helper) are difficult to infer using unsupervised methods, even in closely aligned domains [15]. In addition, as discussed in Section 2.2, HVV has previously been used to study a range of media, including film summaries.

account quality of translation, which weighs into which of two summaries flagged as duplicates are dropped.

[4]The upper bound cut off was also motivated by efficiency constraints for the LLM extraction process.

[5]https://spacy.io/

[6]https://github.com/richardpaulhudson/coreferee

[7]In terms of tags identified by spaCy we include the categories 'PERSON', indicating an individual, 'ORG' an organisation, 'NORP' an ethnic/political group, and 'GPE' a geopolitical entity.

[8]The prompt used for character role label extraction, adapted from Hobson et al. [6], can be found in our code repository.

3.2.3. Encoding for comparison

To enable comparison and evaluation, we use four sentence transformers to encode all summaries (English translations and their anonymised versions) and character roles: sentence-t5-large,[9] bge-base-en-v1.5,[10] nomic-embed-text-v1.5,[11] and snowflake-arctic-embed-m.[12] The first was the best performing model in [5], which we include in order to validate our reimplementation of their work and as a benchmark for the performance of the other models. The other three are currently popular and performant medium-sized general purpose language models,[13] highly ranked on the MTEB leaderboard for English.[14]

We encode the character role labels by passing a single string per summary to the sentence transformers, resulting in one character role embedding for each summary. The string lists the taxonomy-free labels of all the characters, separated under corresponding headings of Hero(es), Villain(s), Victim(s), and Other — an example is included below, in Figure 3.

To test how structured character roles can complement document-level embeddings, we also obtain a combined representation for each summary, defined as the *element-wise sum* of its anonymised summary and character role embeddings. All embeddings were L2 normalised before summation. We tested several different ways of combining the embedding representations, including concatenating (following [20], who concatenated several embeddings after a dimensionality reduction step) and finding their mean (equivalent to addition, up to a normalisation factor). We opted for addition, as this yielded slightly better results and maintained the same dimensionality.

3.3. Evaluation

Computationally, the distance between stories can measured in different ways: (a) *lexically* — comparing edit distance (e.g. Jaccard similarity or Levenshtein distance) or vocabulary sets (e.g. BoW approaches, TF-IDF), like [4]; (b) *semantically* — representing (parts of) stories in latent space where, e.g., cosine similarity can be used to measure distance between embeddings like [5]; or (c) using hybrids thereof — comparing, for example, set distance between semantic representations of narrative components, or other custom models of narrative features like [14]. Here, we opt for a straightforward embedding-based approach using cosine similarity. Although this comes at the cost of some degree of interpretability, it allows us to compare and combine document-level representations with those from structured character roles, similarly to [20], in a way that could easily be extended to other narrative components. Alongside the embedding representations, we also replicate an entity baseline from [5] and perform TF-IDF cosine similarity using the results of our own entity detection process.

In our experiments, we use *hits at k* (H@K) and *mean reciprocal rank* (MRR) as evaluation metrics. H@K measures the likelihood to find a matching label in the first top k model predictions. In our case, H@1 therefore represents the proportion of narratives for which the most similar belongs to the same film cluster; this is equivalent to *precision at 1* (P@1), the evaluation metric used by [4] and [5]. MRR evaluates how well a system ranks the first relevant document for a set of queries, by taking the mean of the reciprocal ranks for all queries.

4. Results

We report two main experimental results: first, an overview of performance on document-level embeddings, including the effects of summary length; and second, the outcomes of our experiment with using character role labels for similarity measurement.

[9] https://huggingface.co/sentence-transformers/sentence-t5-large
[10] https://huggingface.co/BAAI/bge-base-en-v1.5
[11] https://huggingface.co/nomic-ai/nomic-embed-text-v1.5
[12] https://huggingface.co/Snowflake/snowflake-arctic-embed-m-v2.0
[13] For efficiency and consistency, we opted for models with an embedding dimension of 768.
[14] The MTEB leaderboard can be accessed at https://huggingface.co/spaces/mteb/leaderboard

Table 1
Performance of sentence transformers on the full *Tell Me Again!* [5] test set. The same models are compared on the original summaries (those in English and all translations to English) and their anonymised versions. Entity BoW evaluates H@1 on TD-IDF features based on an entity bag of words per summary.

model	H@1 original	H@1 anonymised
bge-base-en-v1.5	0.876	0.382
nomic-embed-text-v1.5	**0.946**	0.448
sentence-t5-large	0.909	**0.555**
snowflake-arctic-embed-m	0.892	0.447
entity BoW	0.855	-

Table 2
H@1 based on full summary embeddings grouped by length, measured by the number of sentences. We dropped any summaries that did not share a label with at minimum 1 other summary per length bin, resulting in a total of 5197 summaries with mean cluster sizes ranging between 2.11 and 2.31. The model names are abbreviated.

		H@1 original				H@1 anonymised			
length	items	bge	nomic	sentence	snowflake	bge	nomic	sentence	snowflake
1≤n<5	1345	0.868	**0.897**	0.833	0.781	0.474	0.494	**0.543**	0.371
5≤n<10	970	0.949	**0.977**	0.960	0.951	0.533	0.577	**0.669**	0.537
10≤n<20	1039	0.950	**0.993**	0.971	0.988	0.501	0.614	**0.724**	0.643
20≤n<40	1600	0.948	**0.995**	0.977	0.994	0.534	0.642	**0.760**	0.740
40≤n<60	190	0.953	**0.995**	0.989	0.989	0.726	0.842	**0.884**	0.868
60≤n	53	0.943	0.868	0.962	**1.000**	0.698	0.566	**0.792**	**0.792**

4.1. Document embeddings

Table 1 gives an overview of the performance of our selected encoders on all summary texts in the test set. In agreement with [5], all models perform significantly worse on anonymised versions of summaries, also outperformed by the entity bag of words baseline (H@1=0.855). Taking a closer look at performance on different length summaries, we find that narrative similarity detection based on document-level embeddings seems to work more reliably — yield higher H@1 scores — when we focus on summaries that are closer in length. As illustrated in Table 2, we find this to be especially pronounced for anonymised summaries.

An interpretation of both findings may be that document-level encoders have difficulty capturing commonalities in narratives beyond surface-level features like character names (illustrated by the high performance on the entity baseline). We expect that a system capable of comparing narrative features, like a shared core plotline or overlapping character set, would be able to abstract commonalities, e.g., canonical events or characters, despite the level of detail in which these are elaborated (granularity associated with the length of a summary). Other, potentially conflating, reasons to consider may also be linked to the anonymisation process, which tends to be poorer on longer texts [5]. Nevertheless, as we see a similar (albeit less pronounced) trend of improved performance on texts of similar lengths for the non-anonymised summaries, we suspect that issues due to anonymisation are not the sole cause for discrepancies in performance when summaries are grouped by length.

4.2. Structured character role labels for story comparison

We find that retrieval based on character role labels alone yields a mean H@1 of 0.222 and H@10 of 0.407 — see Table 3. Given that these comparisons are based on relatively little information — see Figure 3 for an example where the corresponding summaries were 25 (left) and 30 (right) sentences long — we see this as offering a promising first step for a potentially more succinct and structured narrative representation of a full text summary.

Figure 3: Example of character role extraction results for two summaries of Solaris (English translations from Italian and Spanish). Bracketed in grey are the entity names used to query gemma3:4b. Only the text in black was used to generate the character role embeddings.

Table 3
Performance on the narrative similarity detection task using embeddings from anonymised summaries, LLM-extracted character roles, and both combined. The scores are based on a total of 6608 summaries with number of sentences between 5≤n<60. The model names are abbreviated.

	summaries				character roles				combined			
model	H@1	H@5	H@10	MRR	H@1	H@5	H@10	MRR	H@1	H@5	H@10	MRR
bge	0.403	0.567	0.627	0.481	0.211	0.324	0.378	0.269	0.485	0.629	0.683	0.554
nomic	0.495	0.669	0.732	0.575	**0.237**	**0.367**	**0.424**	**0.301**	0.604	0.758	0.809	0.675
sentence	**0.610**	**0.765**	**0.817**	**0.682**	0.213	0.348	0.412	0.281	**0.634**	**0.782**	**0.831**	**0.704**
snowflake	0.538	0.688	0.740	0.609	0.227	0.353	0.413	0.291	0.603	0.744	0.793	0.669

Table 4
Performance on the narrative similarity detection task using taxonomy-free labels *only* to generate character role embeddings (i.e. no HVV information included, in contrast to the results above, in Table 3).

	taxonomy-free character roles				combined			
model	H@1	H@5	H@10	MRR	H@1	H@5	H@10	MRR
bge	0.192	0.325	0.389	0.259	0.485	0.651	0.710	0.563
nomic	**0.270**	**0.416**	**0.483**	**0.343**	0.610	0.767	0.817	0.682
sentence	0.244	0.383	0.449	0.315	**0.626**	**0.772**	**0.819**	**0.694**
snowflake	0.242	0.374	0.434	0.308	0.590	0.732	0.778	0.657

To assess whether character role information can complement document-level embeddings, we also compared performance on summary-only and combined embeddings. As mentioned above, we focus only on the anonymised summaries. We assume these to be a better analogue for "in the wild" applications of narrative comparison (these might include personal accounts, lived experiences, anecdotal evidence, etc.) where we expect to encounter and compare narrative texts that feature characters with different names. Although in some cases the increases are small, we do see an improvement across all metrics summarised in Table 3 when document embeddings are combined with character role representations of the summaries. Among the H@1 scores for example, the average improvement is 0.0699, in relative terms +14.6%; the highest is 0.109, +22.0% (for nomic-embed-text-v1.5) and the best overall performance reaches 0.634 (with sentence-t5-large). A Wilcoxon signed-rank test shows that the changes in H@1 and MRR are statistically significant for all four models ($p<0.001$).

Finally, Table 4 compares performance on the similarity detection task using a slightly different character role representation. The example of Solaris, shown in Figure 3, demonstrates the difficulty in

assigning HVV labels — e.g., in determining whether a character can be considered a hero or a villain — even though the free form labels for corresponding characters might be similar. For this reason, we also tried using a flat character role representation for each summary, based only on the taxonomy-free character role labels. We find that this method performs better when using only the character role data for similarity detection, however is still marginally outperformed by our two-level model in the combined setup.

5. Conclusion and discussion

Although text encoders are powerful tools that enable comparison at a semantic level, our findings suggest that they are (yet) perfectible when it comes to capturing narrative features in text. Our results demonstrate the potential of using character role information for narrative similarity detection, yielding an average improvement in accuracy of 14.6%. For applications that require comparison along particular narrative dimensions, such composite models may be particularly useful.

Nevertheless, as this study was limited to a static representation of character role features, further research using more (and possibly other) narrative schemas is needed to evaluate the effects of combining various components. For character roles, this may include a more fine-grained closed taxonomy (e.g. extending the HVV framework to include other relevant narrative functions, or adopting a more domain-specific one like character tropes in film [21]), or a temporal dimension (e.g. modelling journeys, paths, or arcs that can account for changes in characters' roles over the course of a story). In terms of integrating other narrative features, future work could consider, for example, elements of the plot by proxy of structured (canonical) events, which may prove especially helpful when working with film summaries. If shifting to different story data, such as personal stories or lived experiences, altogether different approaches may be relevant, building e.g. on paradigms like small stories research [26], which may offer concepts better suited to study everyday, mundane, conversational, or unfolding stories.

Another open challenge is in piecing together representations of different narrative features. Future work could therefore experiment further and more systematically with fusion strategies to combine narrative components. In the present study, this was limited to a simple element-wise addition of normalised embeddings. In contrast to the summary embeddings, the character role embeddings were relatively sparse, meaning that in their combined result some signals may be under- or overused. Further investigation involving, e.g., weighted combination, learned fusion, or late-fusion rank aggregation, is needed to better understand how to combine different elements both effectively and meaningfully. Robust, modular fusion strategies will become especially relevant if we want to extend this approach to combining several different structured features.

Finally, we find it important for future work to include the dimension of human perception, building on the cognitivist and participatory approaches that others have employed for narrative comparison [11, 13]. Integrating human judgement into the validation and evaluation process would enable an investigation of (entity and role) extraction quality, perceived similarity, and generalisation to other narrative types. As demonstrated in [6], this may prove particularly helpful in navigating — and perhaps also narrowing — open-ended (taxonomy-free) approaches to labelling narrative components enabled by LLMs. Relatedly, it may also be relevant to consider different evaluation metrics, perhaps some that are more easily interpretable and explainable, such as the character matching approach by [4], or the weighted salience vectors by [14].

Acknowledgments

The authors thank Emiel Verhoef, whose master thesis research was informative for this study.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. C. Schank, Tell me a story: A new look at real and artificial memory., Charles Scribner's Sons, 1990.

[2] B. F. Keith Norambuena, T. Mitra, Narrative maps: An algorithmic approach to represent and extract information narratives, Proceedings of the ACM on Human-Computer Interaction 4 (2021) 1–33.

[3] O.-J. Lee, J. J. Jung, Explainable movie recommendation systems by using story-based similarity., in: Proceedings of the ACM IUI 2018 Workshops, ACM IUI, 2018.

[4] S. Chaturvedi, S. Srivastava, D. Roth, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 673–678.

[5] H. O. Hatzel, C. Biemann, Tell Me Again! a Large-Scale Dataset of Multiple Summaries for the Same Story, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 15732–15741.

[6] D. G. Hobson, D. Ruths, A. Piper, Evaluating Taxonomy Free Character Role Labeling (TF-CRL) in News Stories using Large Language Models, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 14828–14850.

[7] A. Piper, Computational Narrative Understanding: A Big Picture Analysis, in: Proceedings of the Big Picture Workshop, Association for Computational Linguistics, Singapore, Singapore, 2023, pp. 28–39.

[8] V. Propp, Morphology of the Folktale, University of Texas press, 1968.

[9] A. J. Greimas, R. Schleifer, Structural semantics: an attempt at a method, University of Nebraska Press, Lincoln London, 1983.

[10] M. Bal, C. v. Boheemen, Narratology: introduction to the theory of narrative, 3rd ed ed., University of Toronto Press, Toronto, 2009.

[11] B. Fisseni, B. Löwe, Which dimensions of narratives are relevant for human judgments of story equivalence?, in: M. A. Finlayson (Ed.), Proceedings of Computational Models of Narrative, Istanbul, 2012, pp. 114–118.

[12] B. Fisseni, A. Kurji, D. Sarikaya, M. Viehstädt, Story Comparisons: Evidence from Film Reviews, OASIcs, Volume 32, CMN 2013 32 (2013) 94–99.

[13] D. Nguyen, D. Trieschnigg, M. Theune, Using crowdsourcing to investigate perception of narrative similarity, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014, pp. 321–330.

[14] R. Farrell, M. Fisher, S. G. Ware, Salience vectors for measuring distance between stories, in: Proceedings of the 18th AAAI international conference on Artificial Intelligence and Interactive Digital Entertainment, 2022, pp. 95–104.

[15] L. Jahan, R. Mittal, M. Finlayson, Inducing stereotypical character roles from plot structure, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 492–497.

[16] P. Gervás, G. Méndez, Tagging narrative with propp's character functions using large language models., in: Proceedings of the Text2Story'24 Workshop: Text2Story @ ECIR, CEUR-WS, Glasgow, United Kingdom, 2024, pp. 137–148.

[17] K. Gehring, M. Grigoletto, Analyzing climate change policy narratives with the character-role narrative framework, SSRN (2023).

[18] S. Fharook, S. S. Ahmed, G. Rithika, S. S. Budde, S. Saumya, S. Biradar, Are you a hero or a villain? a semantic role labelling approach for detecting harmful memes., in: Proceedings of the workshop

on combating online hostile posts in regional languages during emergency situations, 2022, pp. 19–23.
[19] D. Stammbach, M. Antoniak, E. Ash, Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data, in: E. Clark, F. Brahman, M. Iyyer (Eds.), Proceedings of the 4th Workshop of Narrative Understanding (WNU2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 47–56.
[20] J. Elfes, Mapping News Narratives Using LLMs and Narrative-Structured Text Embeddings, 2024.
[21] D. Bamman, B. O'Connor, N. A. Smith, Learning latent personas of film characters, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 352–361.
[22] E. Chu, P. Vijayaraghavan, D. Roy, Learning personas from dialogue with attentive memory networks, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2638–2646.
[23] A. Piper, S. Bagga, Using Large Language Models for Understanding Narrative Discourse, in: Proceedings of the The 6th Workshop on Narrative Understanding, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 37–46.
[24] L. Jahan, R. Mittal, W. V. Yarlott, M. Finlayson, A straightforward approach to narratologically grounded character identification, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6089–6100.
[25] G. Team, Gemma 3 (2025).
[26] A. Georgakopoulou, Small stories research: Methods–analysis–outreach, The handbook of narrative analysis (2015) 255–271.